Internal Report 2012-19

Universiteit Leiden
Opleiding Informatica

Ready for the future in health informatics: towards semantic interoperability in dermatology

L.I. van der Meer

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

BACHELOR THESIS
LUCAS VAN DER MEER

"Ready for the future in health informatics:
towards semantic interoperability
in dermatology"

Version:	2.1 (Final)
Date	Friday, September 28, 2012
Supervisor:	Fons J. Verbeek (LIACS)
Second reader:	Maarten Lamers (LIACS)

Leiden University
Leiden Institute of Advanced Computer Science

ABSTRACT

"Ready for the future in health informatics:
towards semantic interoperability
in dermatology"

Summary: Health Informatics applies the advancements of computer science into the medical domain. A requirement for the achievement of such advancements is reaching a certain level of semantic interoperability.

This thesis describes how semantic interoperability can be achieved at the dermatology department of the Leiden University Medical Centre (LUMC), resulting in better content-based retrieval of images, improved ability to gather statistics, and computer-aided diagnosis & telemedicine.

Techniques that are used to accomplish these benefits are including ontologies, image annotation, standards, Machine Learning, Computer Vision, Information Extraction and Data Mining.

A list of recommendations to accomplish semantic knowledge integration, and an example interface that supports the process of manual image annotation, are given.

Conclusions and recommendations: To achieve the described benefits, the LUMC and the dermatology department should reach a certain acceptable level of semantic interoperability. This can be done by using a decent, computer-processable code system, by annotating data and by using standards for information exchange and for imaging.

Keywords: Semantic interoperability, information-as-knowledge, data annotation, ontologies, standards, Machine Learning, Computer Vision, Information Extraction, Data Mining, Dermatology, Health Informatics.

TABLE OF CONTENTS

1. INTRODUCTION

Chapter 1 introduces semantic interoperability and explains why integrating a decent computer-processable nomenclature, data annotation, and standards for information exchange and imaging into the medical practice contribute to achieving it. Chapter 2 discusses all the techniques in detail, finally chapter 3 recommends how the LUMC dermatology department can integrate all techniques into workable solutions.

1.1 Challenge statement

The field dealing with how the medical domain can benefit from computer science is called Health Informatics[1]. A requirement for many of these benefits is that the IT systems that are used achieve a certain acceptable level of semantic interoperability. The challenge is to integrate the necessary changes into the medical practice. Where changes in the workflow of the staff are inevitable, the key is to convince the staff that these changes will eventually result in benefits. This requires a close cooperation between medical staff and computer scientists.

With the emergence of computer systems in healthcare come challenges for semantic interoperability: patient information is stored in Electronic Health Records (EHR), images of MRI scanners are processed by computers and digital images of patient disorders are produced. These are all different systems producing a varied, huge[a] amount of data. Making all these systems communicate effectively with each other and ensuring that one is able to retrieve information is a major challenge.

Semantic interoperability is a requirement for many other techniques from computer science to succeed. Only then, hospitals can fully benefit by the better retrieval of information by their content, by performing computer-aided diagnosis and telemedicine, and by the improved ability to gather statistics about disorders; all of which will eventually result in better healthcare, education and research (Korenblum *et al.*, 2011)[2].

Figure 1 illustrates the necessary steps to take advantage of computer science in the medical domain: after integrating some basic techniques into the medical practice, semantic interoperability can be achieved after which, may or may not via other informatics techniques, benefits will result.

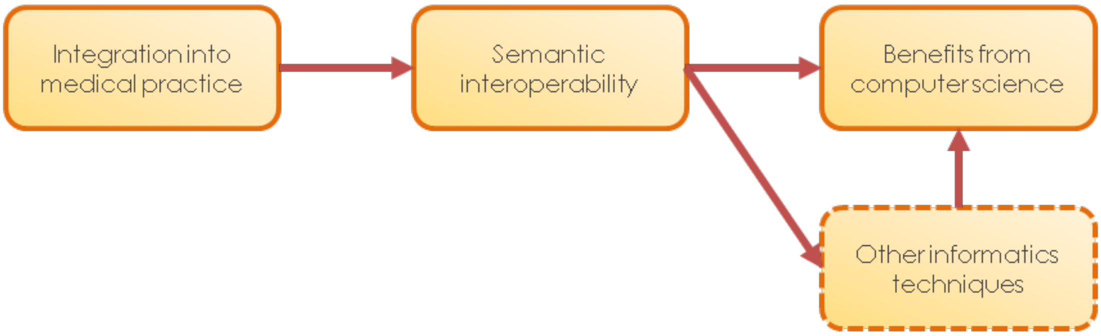


Figure 1
Taking advantages of computer science in the medical domain by achieving semantic interoperability.

[a] Of course, 'huge' is relative. It is not huge compared to the data produced at for instance LOFAR (*www.lofar.org*), but still huge enough to consider it as a major challenge for interoperability.

Semantic interoperability

The techniques to achieve semantic interoperability are by using a decent, computer-processable nomenclature, by annotating data, and by using standards for information exchange and imaging. These techniques will be referred to as the basic techniques.

Other informatics techniques

Often semantic interoperability cannot directly result in benefits for healthcare, therefore it needs other techniques such as Information Extraction, Computer Vision, Data Mining and Machine Learning. These techniques will be called the supportive techniques.

Cooperation

Health Informatics operates on the intersection of computer science and medical science. Integrating all suggestions from computer scientists into the medical practice is a difficult task requiring a close cooperation between these disciplines.

The domain

This thesis aims specifically at advising how to achieve semantic interoperability for the handling of images at the dermatology department of the Leiden University Medical Centre in the Netherlands.

Ready for the future

The IT systems that are currently used in the hospital and the department have cost a lot of money and many people are working with them on a daily basis; replacing them as a result of any paper thus seems unlikely. Also, some of the techniques that will be used have potential but are not yet mature. Therefore the goal of this article is to recommend what this department needs to do in order to be ready for the future in health informatics; hopefully will this thesis encourage people to reflect on their current systems and will it motivate them to put time into achieving semantic interoperability: the efforts will eventually be worth it.

1.2 Semantic interoperability

The term interoperability explicitly mentions the ability to reuse data[3,4,5]. Every application of computer science depends on such interoperability, hence the term information technology (IT): the branch of engineering dealing with the use of computers to store, retrieve and transmit information[6]. A loose interpretation of IT is *the branch of engineering dealing with interoperability in computer systems.* Interoperability is a requirement for many informatics technique to succeed. Semantic interoperability is reusing data in such a way that the meaning of it is unambiguous and shared.

Semantic interoperability

- Annotated data
- Standard nomenclature
- Standards for information exchange and

Achieving semantic interoperability in digital systems is done by integrating semantic knowledge: the integration of data with the *meaning* of it. In other words, machines have to be learned what exactly data mean. Three ways to do so are covered: by annotating data, by using a standard nomenclature, and by using standards for information exchange and imaging.

1.2.1 Annotated data

One way to do this is by annotations. How annotations are used to facilitate semantic knowledge integration can be illustrated with the following example:

> Imagine that you have an image collection of personal photographs. If you would like to search for images of all your birthday parties, this is not possible when the images are 'bare', i.e. contain no extra information. Luckily, all modern cameras automatically store the date and time that the photograph is taken in the same file as where the actual pixels are stored. These are data (date, time) about the content of data (the actual photograph; the pixels): metadata, or annotations. As all modern computers understand these codes, you are able to query for all photographs ever taken on the 12th October and thus find the images of all your birthday parties.

The computer 'learned' that October 12 is a date, and that October 12, 2000 and that same date in 2001 belong to each other: it learned about the meaning of data. Note that annotations do not have to relate to images, they can be used to describe data about any data content.

Direct image annotation

Images are one form of data that can benefit the most from annotation as their semantic content cannot directly be searched for. A distinction is made between direct annotation in which information is directly added to the images (for instance when drawing a circle to mark the wound area; Figure 4) and indirect annotation (for instance adding the diagnose of a particular disorder to an image).

Figure 2 - Making the lesion area visible (direct annotation)

The dermatology department of the LUMC uses Photoshop when directly annotating images. The tool that manages all images is not used for that purpose as the tool is found inconvenient by its users. As a result, the annotations are not stored as for instance an XML file, but are embedded in the images. This limits the reusability of such annotations.

Another way of direct annotation is physically annotating the object that is being photographed, for instance by drawing an annotation directly on the skin, or by holding a measuring tape to the skin to indicate the size of the picture area. An example of both can be seen in Figure 3. Physical annotation is common in the dermatological domain. From a computer science point of view this is undesirable as physical annotations modify the skin thus eliminating information. A more desirable way is to digitally mark the lesion area or measure the size of the object that has been photographed. The latter can be done using image processing and analysis; to do so other variables have to be taken into account, such as the orientation of the photographed object and the amount of pixels in the image.

Figure 3 - Physical annotation

Naturally, annotating images takes time and effort and the results of it may not directly be visible. Thus the challenge is to integrate the direct annotation of images into the routine workflow of the staff.

The next step in achieving semantic interoperability is the use of standards: agreements made by a large amount of people and institutions about what an event or instance should look like.

1.2.2 Standard nomenclature and ontologies

When looking at nomenclature, or terminology, standardisation means making agreements about the meaning terms and about how do they relate to each other. This decreases error by miscommunication, between humans and between computer systems.

The relationship between objects is defined by ontologies. This way a computer knows how objects are hierarchically ordered, hence integrating data with its meaning.

1.2.3 Standards for information exchange and imaging

From a data perspective, standardisation means agreeing on how data are produced, stored, presented and transferred.

1.3 Benefits from computer science

The advancements of semantic knowledge integration are illustrated using three example situations representing three different perspectives.

Benefits from computer science

- Content-based retrieval of images
- Improved ability to gather statistics
- Computer-aided diagnosis

1.3.1 Content-based retrieval of images

Medical researcher:

> "I want to find all images having an instance of *Spinocellulaircarcinoom* that occurs on the hands"

To allow this query, the 'regular' retrieval of digital data – browsing through a list of files – is insufficient as images cannot directly be searched for its content. Storing additional information in the form of annotations about images, such as the object and the disorder that are photographed, allows searching for the content of images. This is further facilitated by storing the meaning of the annotations, using ontologies. This allows a computer to understand that a photograph that is for instance annotated as *primary_limb = knuckle* should be taken into account in a query for hands, as knuckles are part of the hand.

Standard terminology for medical terms eliminates ambiguity thus further facilitating image retrieval based on its content. The same goes for standards for information exchange, for taking images and for storing annotations as they ease the exchange of information between systems.

1.3.2 Gathering statistics

Public health investigator:

> "Ten years ago, legacies that allow airplanes flying to Schiphol Airport to discharge kerosene right before landing, were introduced. I want to investigate if this has consequences for the health of people living in the area where kerosene is often discharged[b]."

This researchers needs a lot of data and will ask the LUMC to provide them. These data may not be connected, such as disorder history, the time that certain disorders occurred and the living place at a certain time. There may be a correlation between the occurrence of disorders as a result of the changed legacies and skin complexion, age or sex. Using techniques from computer science, many information can be integrated making it easier to gather such data.

[b] This is a fictive scenario.

1.3.3 Computer-Aided Diagnosis & Telemedicine

Computer scientist:

> "In order to apply computer-aided diagnosis on images, I need to train a classifier that can distinguish disorders that occur in these images. I want to do so by applying Machine Learning on a large set of medical images"

Computer Vision (using Machine Learning) has the potential to allow computers to make diagnoses. To train a classifier, the computer scientists needs a large set of photographs that are annotated with the disorder. Other information, such as skin complexion, the primary body part shown in the picture, age or sex may also be needed. The lesion area should also be annotated in at least part of the collection. As computer scientists are not allowed to see pictures containing genitalia or recognisable features such as tattoos, a feature that can easily hide these images is desirable.

Advancements in personal computing, digital cameras and smart phones allow individuals to easily send a digital photo to a hospital where a physician is able to look at it; advancements in the presence of the Personal Health Record also contribute to Telemedicine.

1.4 Domain analysis: a hospital's dermatology department

A proper advice cannot be given without knowing the domain in which the advice will be given. A brief analysis of the domain is made: the Dutch *Leiden University Medical Centre* (LUMC), its dermatology department, the IT systems that are used, and in general the challenges of working with IT systems in a hospital.

1.4.1 The LUMC and its dermatology department

The *Leids Universitair Medisch Centrum* is a university hospital located in Leiden, the Netherlands, affiliated with Leiden University. It has around 7.000 employees taking care of 20.000 clinical hospitalisations each year. A university hospital means that there are also students being educated and that research is being done at the hospital. Annually, 2.500 students are being educated and over 200 papers are published[7]. Its dermatology department is relatively small, having about 50 staff members. They give 30.000 consults each year[8].

1.4.2 Research

A university hospital means that researchers are working in it. Researchers require more from data than physicians do, adding extra challenges to how the collection and processing of data is taking care of in the medical practice.

1.4.3 Quality care

Working in a hospital means that the very first priority is the quality of healthcare. This statement implies a risk as well as an opportunity. On one hand, working on any system in a hospital means working directly on the provision of healthcare, therefore changes to the systems should not disrupt this healthcare. On the other hand, the ongoing quest for quality care can be a great incentive to evolve to better systems. When one is capable of implementing new information systems while patient care is not disrupted, great improvements in the quality of that patient care can be accomplished.

1.4.4 Staff

One additional challenge lies in the fact that hospitals are dynamical environments in which the working pressure is high; as a result the staff can put little time into tasks that do not belong to their primary ones, such as the implementation of new IT systems. Also, various IT systems have been implemented in the past of which the staff might not be so enthusiastic about. Teaching them new systems or interfering with their workflow might encounter major resistance – even when the staff are convinced that they, as well as the patients, will eventually benefit from these changes. Keeping the users of the systems and their concerns constantly in mind is crucial to the success of any change.

1.4.5 Dermatology

Dermatology is the branch of medicine dealing with the skin and its diseases. A dermatologist takes care of diseases of the skin, scalp, hair, and nails[9]. A specific jargon is used, of which advantage is taken in section 2.4.3.

As dermatology is a visually based medical disciplines, taking photographs of symptoms is a common procedure.

1.4.6 Health insurance

As on January 1, 2006, every person living in the Netherlands is obligated to have health insurance[10,11].

Even though the basic coverage (mainly ordinary treatments) of health insurance is determined by the government, insurers have a major influence in determining what less-ordinary treatments are covered. When an experimental treatment – such as light therapy for a naevus – has been performed, insurance companies request the treating hospital to send them photographs of the treatment progress. This way they can gather information about whether treatments are effective and therefore whether they should pay for it.

As a result, health insurers have a marginal influence on how IT systems in the medical domain look like. When the recommendations are implemented, this will result in improved communication between these insurers and medical institutions.

1.4.7 Privacy issues

Working in a hospital environment means that a large part of the data are confidential. This causes problems when non-doctors such as computer scientists are working with these data.

Non-doctors must sign an agreement in which they promise to handle all information with care. A person who has signed this agreement is allowed to see and work with a selection of the photographs: excluded from this selection are all pictures containing recognisable parts (faces, tattoos, wedding rings), or genitalia (including women's breasts). Non-doctors can never work with other patient data such as name, address and treatment history.

The system that is used to manage the image files has no feature to hide all classified data. This means that when non-doctors need to work with the system, doctors manually need to construct a selection of appropriate images. As this is a time consuming task, a special mode to display only certain, approved, photographs would be greatly appreciated.

1.4.8 IT systems

Various automated systems are being used on a large scale, both in software and hardware applications. Examples are the Electronic Health Record, Hospital Information System and equipment for CT and ECG scans. The majority of such systems is developed and maintained by private organisations and implementing such a system may cost as much as several hundred million Euro's.

When an organisation or a department needs a new IT system, the common first step is to ask the highly specialised medical staff of that division to list the requirements of the new system. As a result the main focus usually lies on the specific needs for that division and less on interoperability. Consequently, an enormous amount of various systems and standards are being used in the medical domain. In the Netherlands, there is little alignment to any system or standard, resulting in challenges to information interoperability. It is no exception that within a hospital, or even within a department, various, competing systems and standards are being used (Peter Hendriks, 2012; unpublished work)[12].

In general there are two major systems that are being used in hospitals: the Hospital Information System (HIS; including the Electronic Health Record and Diagnose-Behandelcombinatie), and Picture Archiving and Communication Systems (PACS), used to manage image files. Figure 4 shows a representation of the biggest IT systems in a hospital:

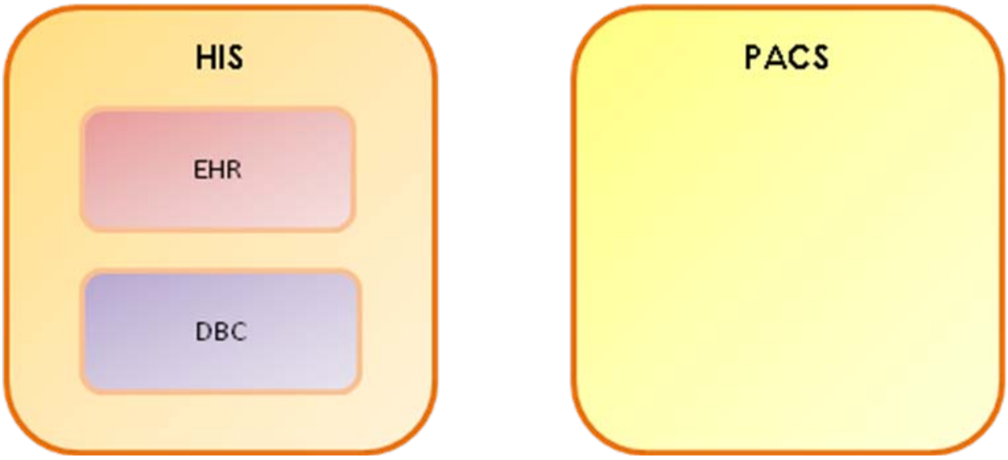


Figure 4
A representation of biggest IT systems in a hospital

Hospital Information System

The most prominent system in a hospital is the Hospital Information System (HIS), or *Ziekenhuisinformatiesysteem (ZIS)* in Dutch. The term HIS will be used throughout this thesis. A HIS essentially is a computer system that is designed to manage and integrate all digital information to support staff in doing their jobs effectively. The minimal coverage of such a system is administration, support for logistics and for the financial handling of treatments[13]. In practice a HIS contains only a portion of the available digital information, stretching the need for standards to ensure effective communication between different systems. Many different HIS's are available on the market.

The HIS used by the LUMC is eZIS, developed by Chipsoft[c]. The EHR and DBC (both discussed further on in this session) are incorporated in eZIS. The system was introduced in 2011; before several separate systems were used.

Electronic Health Record

An Electronic Health Record (EHR), or *Elektronisch Patiëntendossier (EPD)* in Dutch, is a software application where medical records are digitally stored in and can be retrieved from. EHR's are designed to integrate patient data from various disciplines. Most EHR's are included in a HIS.

In the Netherlands, a nationwide accessible EHR was introduced on November 1, 2008; its main target is to reduce errors in the exchange of information between medical specialists. On April 5, 2011, the Senate (*Eerste Kamer*) rejected this nationwide accessible EHR because there were "too many risks for reliability, safety and privacy"[14]. In September 2011, the National IT Institute for Healthcare in the Netherlands (Nictiz[d]) decided, with the support of the minister of Public Health, to relaunch the project using private funds. After immediate objections by the Lower House (*Tweede Kamer*), the project was cancelled only a month after its relaunch. Currently, new attempts to introduce the nationwide accessible EHR are made by the private sector, but not under the supervision of the minister of Public Health.

For the time being, medical institutions in the Netherlands are using the Electronic Health Record of their own choice.

---

[c] www.chipsoft.nl

[d] www.nictiz.nl/page/Over-Nictiz/About-Nictiz

Personal Health Record

A different form of the Electronic Health Record is the Personal Health Record (PHR), in which patients are able to glance at their medical information themselves.

Nomenclature

The strength of EHR's lies in their usage of terminologies decreasing errors by miscommunication. When explicit codes are connected to terms, terminologies are often called *code systems*. When terminologies are hierarchically ordered, they are usually called taxonomies, thesauri or ontologies. Most commonly used are ontologies. Possible applications of ontologies in the healthcare domain are the hierarchy of diseases and of body parts. Using ontologies, better information storage and retrieval can be accomplished.

There are two major code systems used in the Dutch medical world: SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) and ICD (International Classification of Diseases). The LUMC uses the Clinical Modification of the latter, version 9 (ICD-9 CM).

It is important that physicians do not have to work with codes but can use their own terminology; the system is responsible for translating these terms into explicit codes. All codes should have an unambiguous meaning.

Diagnose-Behandelcombinatie

Diagnose-Behandelcombinatie (DBC), or Casemix in English, is a code system that is being used in the Dutch healthcare system to describe a delivered care product. As DBC is a crucial term in the Dutch care system, this Dutch term will be used throughout this thesis. DBC's are often implemented in HIS's.

A DBC is a representation of all activities and operations in a hospital that *one single* patient runs through during a period of time, related to *exactly one* complaint. It was introduced in January 2005 to measure hospital performance, aiming to reward initiatives that increase efficiency. The reasoning behind this DBC is that when hospitals can only receive honorarium for completed DBC's, they are motivated to work efficiently.

Picture Archiving and Communication System

Picture Archiving and Communication Systems (PACS) are systems that are used in the medical world to store, retrieve and transmit digital images from multiple modalities[15], such as radiology and dermatology. The universal standard used by PACS to store and transfer images is DICOM (Digital Imaging and Communications in Medicine).

The PACS used in throughout the LUMC is called Clinical Assistant (CA)[e]. A feature in eZIS makes it possible to look at photographs that are stored in CA using the eZIS interface. The dermatology department uses CA to store photographs of patients' disorders.

1.4.9 Dissatisfaction about the IT systems

Among the LUMC staff, there is some dissatisfaction about the IT systems that are being used. In general, the LUMC has the ambition of exploring new, modern ways to communicate with patients, but has the feeling that the IT systems do not support them in doing so (in fact, the staff feels that they are held back by the systems).

The opinions in this paragraph are collected by talking to one doctor and one medical photographer in their working environment; the goal of the conversation was to get a complete view of the IT systems in the hospital. The two persons

[e] www.rvc.nl/EN/Clinical-Assistant

that the author has talked with have collected many opinions from other staff, therefore their statements are believed to be representative. Further research using a larger, more diverse group of staff would clarify if this indeed is the case.

Nomenclature

The LUMC uses ICD-9 as its terminology system. Because no dermatologists have been involved in the development of this standard, it has limited opportunities to describe diagnosis and treatment. As a result, the dermatology department altered the standards for internal use.

Electronic Health Record (in eZIS)

The EHR (as part of the HIS eZIS) is the most commonly used system by physicians, but they are unhappy with it. The following are the most important complaints:

- *"Relevant information is difficult to find"*

  It is common that a hospital patient visits several doctors. Every doctor examines the patient and writes a report in eZIS. Such a report is extensive; one A4 of plain text is no exception. When the next doctor examines the patient it first needs to find relevant information about the session with the previous doctor, such as the diagnose and medication; any other information often is irrelevant.

  As gathering information is a key aspect of doing research, the lack of information retrieval is particularly affecting the researchers at the hospital.

- "The way to input my information is inconvenient"

  eZIS tries to structure the way that the staff add information into the system, targeting better information retrieval. One common way of doing so is by spreading information over multiple input fields, such as 'observations', 'possible disorders' and 'possible medication'. During an examination it is difficult to split up these activities, resulting in a disruption of the workflow of physicians. Most physicians are not unwilling to structure their input, but the disruption results in irritation and possibly also in worse information retrieval because staff tend to add information into the same input field.

Diagnose-Behandelcombinatie (in eZIS)

The downside of the DBC system is that when a disorder occurs more than once on a patient in the adjusted period of time, only one of the disorders can be charged for. This happens far more often than average in the dermatological domain and means a serious financial disadvantage to the department.

DBC offers limited functionality in the naming of disorders, diagnoses and treatments; it is far less advanced than the (not so advanced) ICD. An often heard statement by physicians is "I am not putting effort in neatly describing my medical terms, because when they enter the DBC they end up in the same lump." Thus, this limited functionality demotivates the willingness to use the proper terminology, hence decreasing the possibilities for semantic information retrieval.

Integration eZIS & Clinical Assistant

All images are stored in CA; when a doctor is treating a patient he or she wants to see these photographic images. A plug-in in eZIS makes it possible to browse for all photographs that are stored in CA, but this structure has its limitations.

eZIS is a patient-oriented system, whereas CA is diagnose-oriented. As a result, in eZIS it is only possible to show photographs that are related to the patient file one is working in; it is for instance not possible to browse for all pictures containing one particular disease.

As Clinical Assistant is a diagnose-oriented system it is practically useless unless patients diagnoses are available in it. When a doctor updates a patient's file in its EHR (eZIS), the connection constraint with CA prevents the information from automatically being transferred to CA. Therefore this has to be done manually.

Clinical Assistant

CA has limited opportunity to store image annotations. The department stores what body parts are shown in the column 'remarks', but only when the staff thinks that it is too hard to determine the body parts manually.

A feature to directly annotate images is available in CA, but the tool is hardly being used because the medical staff reasons that "most humans are able to see what the lesion area is, so I will not put time into marking it". Also, the feature is found inconvenient to use. Therefore, other programs such as Photoshop are used for direct annotation. Consequently, the original photograph is only indirectly available using the Photoshop file.

Also, doctors and researchers would like to search for people with a specific skin tone because there is a strong correlation between skin tone and certain diseases. Such a feature is unavailable.

Most departments of the LUMC use DICOM when working with photographic images, others including the dermatology department do not. The main reason for this is that specific requirements for dermatologists, such as guidelines for working with polarisation and wavelength filtering, are not met in DICOM. Also the absence of specialised machines such as those producing for X-Ray images takes away the necessity to use it. The interested reader is suggested to read the latest paper draft by the working group for dermatologic standards in DICOM ("working group 19")[16].

1.4.10 The process of taking images

As this thesis focuses on the advantages for dermatologic imaging it is important to know how the photo taking process looks like.

When a physician thinks that taking a picture is relevant (e.g. because it is a rare disorder, to determine improvements as a result of treatment, or when an insurance company requests so), he or she directs the patient to the photographer.

The photographic images are usually taken in a professional studio that is located in the department. An equal background is established by placing the patient in front of a bluescreen. A dressing room is present. A single-lens reflex camera (SLR) is used to ensure high image quality. See Figure 5 for a photograph of the image studio.

Figure 5 - The image studio

File and annotation specifications

Images are usually taken using a high quality SLR camera, mainly a Nikon D300 or Fuji Finepix S3. They are shot in RAW format with a resolution of 13.1 Megapixel after which they are edited if needed using Adobe Photoshop[f]. After editing they are being saved in the TIFF file format.

Before the pictures are stored in Clinical Assistant they are converted into JPEG; the TIFF files are stored in an archive.

Not all images in the database are taken this way. Sometimes, when a patient is an infant, or infirm, the photographs are taken elsewhere in the hospital (e.g. the patients bed). Some images are taken at home by the patient itself, by the general practitioner or by another hospital and then sent via e-mail; others are taken using a dermatoscope (a camera device specifically build to examine the skin; Figure 6 and Figure 7). As a result of many different persons taking the photographs using many different camera types, differing image qualities remain.

Figure 6 - A dermatoscope

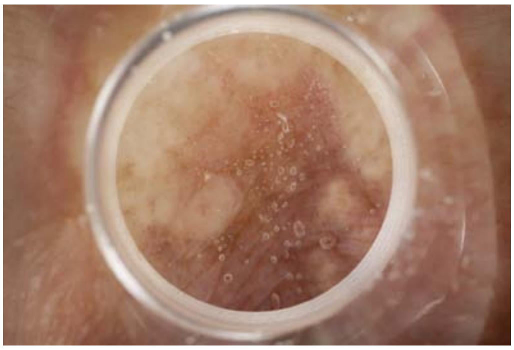


Figure 7 - An image taken using a dermatoscope

[f] *www.adobe.com/Photoshop*

The photographer

The photographer is not a doctor but has some basic knowledge about skin diseases and patient care. It is working at the department four days a week; the salary is paid from research budget. Taking and managing all photographs does not take up all of the photographer's time; the photographer also takes care of some supporting tasks at the department, such as computer help.

Workflow when taking photographs

When a physician wants a photograph to be taken it sends a request to the photographer using the HIS. The photographer then takes the photographic images and enters them into the PACS. Only when the physician encloses the patients' diagnose with the request, this diagnose is directly added into the PACS. Otherwise, it is added later by a department secretary.

The doctor keeps track of the patient information (for instance updates on the curing process) using the HIS; the photographer uses a different database (the PACS) to upload the photographic images to. Both databases are limitedly connected.

1.5 Research objective

In the following chapters, data annotation, ontologies, standards, Machine Learning, Computer Vision, Information Extraction and Data Mining are highlighted after which recommendations for the implementation are given. Goal is to achieve semantic interoperability, allowing the retrieval of data based on their content, improved ability to gather statistics and computer-aided diagnosis & telemedicine.

As switching to new IT systems often leads to draw-backs and because some of these techniques are not yet ready for the future, the goal of the author is not to persuade the department to embrace all suggestions. Rather, it is to show the potential benefits of informatics techniques and what can be done in order to be ready for them.

1.5.1 Subquestions

Finding the answer to what the LUMC dermatology departments needs to change in order to achieve benefits as a result of semantic knowledge integration, consists of answering several subquestions. At first, the current situation at the department needs to be explored, after which semantic interoperability and accessory techniques need explanation. Thereafter are other techniques to accomplish the desired benefits explained; finally recommendations for the future of the LUMC dermatology department can be given.

The subquestions are:

1. **The current situation at the LUMC dermatology department**
   - 1.a. What information systems are being used, especially related to handling images;
   - 1.b. How is the accomplishment of semantic interoperability being handled;
   - 1.c. How do the staff judge the usability of these systems and techniques;
2. **Techniques to achieve semantic interoperability**
   - 2.a. What is semantic interoperability;
   - 2.b. What are the basic techniques to accomplish semantic interoperability;
   - 2.c. What standard terminologies and code systems are available;
   - 2.d. What tools and techniques are available for data annotation;
   - 2.e. What standards for information exchange and imaging are available;
3. **Supportive techniques**
   - 3.a. What other techniques that support the accomplishment of the desired benefits are available;
   - 3.b. How usable are these techniques;
4. **The future of the LUMC dermatology department**
   - 4.a. How should the basic techniques be handled;
   - 4.b. How can the process of achieving semantic interoperability be supported and automated.

1.5.2 Research area

This research covers the fields of Computer Science (more specifically: Semantic Interoperability, Information management, Requirements Engineering, Data Mining, Computer Vision, Machine Learning) and Medical Sciences (Dermatology, Health Informatics, Computer-Aided Diagnosis, Telemedicine).

1.5.3 Research methodology

In 2010, Kumar[17] defined different research methodologies for several application fields, inquiry types and objectives.

The application field of this research project is applied research: applying theoretical knowledge in order to improve the design of systems. When looking at the inquiry type, Kumar distinguishes between quantitative and qualitative research. This thesis describes the benefits for people of informatics techniques; this can best be guided by a qualitative approach. Kumar's final distinction is involved with the research objective. As most of the research consists of describing the current situation at the department, describing the options to improve the situation, and thereafter giving an advice to accomplish this, the descriptive research methodology will best guide this research.

2. SYSTEMS & STANDARDS

This chapter describes all the systems and standards that can be used to integrate semantic knowledge at the dermatology department of the LUMC. Together with other techniques from computer science, they will result in concrete benefits for the department.

One of the key principles of computer science, semantic interoperability, is explained, and related to that are explained the ambiguity of information and of annotation, and ontologies and standards. Finally, the informatics techniques Machine Learning, Computer Vision, Information Extraction and Data Mining are explored.

The chapter consists of three parts: the first covers semantic interoperability and techniques to achieve it, the second covers other informatics techniques, the third part explores the preliminary conclusions.

PART 1
SEMANTIC INTEROPERABILITY AND ACCESORY TECHNIQUES

Part 1 of this chapter describes semantic interoperability and techniques to achieve it.

2.1 Fundamentals of IT: Semantic interoperability

Taking good care of data – that is: ensuring interoperability – is one of the major problems in Computer Science.

2.1.1 The Semantic Web

The best example of semantic interoperability – and at the same time the ultimate goal to be reached – is the Semantic Web, as named by the founder of the original web, Berners-Lee, in 2001[18]. He explains the Semantic Web with the following example:

> *"The entertainment system was belting out the Beatles' "We Can Work It Out" when the phone rang. When Pete answered, his phone turned the sound down by sending a message to all the other local devices that had a volume control. His sister, Lucy, was on the line from the doctor's office: "Mom needs to see a specialist and then has to have a series of physical therapy sessions. Biweekly or something. I'm going to have my agent set up the appointments." Pete immediately agreed to share the chauffeuring.*
>
> *At the doctor's office, Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved information about Mom's prescribed treatment from the doctor's agent, looked up several lists of providers, and checked for the ones in-plan for Mom's insurance within a 20-mile radius of her home and with a rating of excellent or very good on trusted rating services. It then began trying to find a match between available appointment times (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules. (The emphasised keywords indicate terms whose semantics, or meaning, were defined for the agent through the Semantic Web.)*
>
> *In a few minutes the agent presented them with a plan. Pete didn't like it—University Hospital was all the way across town from Mom's place, and he'd be driving back in the middle of rush hour. He set his own agent to redo the search with stricter preferences about location and time. Lucy's agent, having complete trust in Pete's agent in the context of the present task, automatically assisted by supplying access certificates and shortcuts to the data it had already sorted through.*
>
> *Almost instantly the new plan was presented: a much closer clinic and earlier times—but there were two warning notes. First, Pete would have to reschedule a couple of his less important appointments. He checked what they were—not a problem. The other was something about the insurance company's list*

failing to include this provider under physical therapists: "Service type and insurance plan status securely verified by other means," the agent reassured him. "(Details?)"

Lucy registered her assent at about the same moment Pete was muttering, "Spare me the details," and it was all set. (Of course, Pete couldn't resist the details and later that night had his agent explain how it had found that provider even though it wasn't on the proper list.)"

The Semantic Web is a digital system that understands persons and as a result provides the best solutions. It is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

2.1.2 The ambiguity of information

To work on the fundamental challenge of ensuring semantic interoperability, a clear definition of the term 'data' has to be made. People tend to think of data as information, but considering data as information alone is insufficient as information implies an informative value. When data are being poorly handled (i.e. interoperability is bad) they cannot be retrieved; therefore their informative value cannot be used. Thus, the goals for all data is to be informative: *data = information*.

In his famous 1991 paper, Buckland[19] notes that the term information is used in different ways and he tackles this ambiguity by defining three 'new', ideally distinct, types of information: *information-as-process* is the act of informing; *information-as-thing* are objects, such as data and documents, that are referred to as information because they are regarded as being informative; *information-as-knowledge* is the actual intelligence, knowledge.

The actual intelligence (in the hospital database systems: the *meaning* of photographs and patient information) is information-as-knowledge; the *representation* of it (a photograph stored as for instance a TIFF-file; a patient record in a database) is information-as-thing. Thus, a good connection between these data and the semantic *meaning* of them has to be present.

Ultimately, information systems can deal directly *only* with information-as-thing. The task of such systems can be defined as *dealing with data (information-as-thing) in such a way that it becomes intelligence (information-as-knowledge)*.

The next section describes one way of dealing with data to ensure its informative value: by adding annotations.

2.1.3 The ambiguity of annotations

In chapter 1 the term annotations has been described as 'data about the content of data'. A more exact definition has been made by MacMullen (2005)[20], following Buckland's work: *annotation-as-process* is the act of annotating; *annotation-as-thing* is the note linked to an information object (i.e. the physical presentation of the annotation); *annotation-as-knowledge* is the semantic meaning of an annotation. By using this unambiguous terminology one is able to make exact definitions and compare outcomes across domains.

Many things can be pointed as annotation-as-thing, including a circle drawn on an image; the fact that for instance a person visited the jungle is coupled to the Electronic Health Record; the coupling between 'head' and 'torso'; as well as stored information about a photograph such as the focal distance that was used when taking it.

In this example, adding annotation-as-knowledge to annotation-as-thing consists of explaining a machine that: the circle means an infected area; 'has visited the jungle' means an increased risk of certain diseases; the coupling between 'head' and 'torso' means that they are physically connected in the body; and the focal distance can be used to determine the distance between the photographed object and the camera.

2.1.4 Adding annotation-as-thing to images

Especially images can benefit from having proper annotations as their semantic content cannot be searched for directly. Web-based systems to annotate images on the web have already been developed, such as LabelMe[21].

In the medical field attempts have been made to allow medical staff to easily annotate images as part of their routine workflow. Rubin *et al.* (2007) developed AIM[22]: a system to store and retrieve annotations of medical images; the structure of the annotations is guided by an ontology. One year later he and others built *i*Pad[g,23]: an interface tool that guides the annotation process of doctors. It uses AIM as a backbone. It has a controlled vocabulary to ensure that the medical staff use the appropriate terminology. In the succeeding paper by the same research group, BIMM**2** was presented: a system that searches for and retrieves images that are stored by using AIM and *i*Pad (Korenblum *et al.*, 2011). A different research group built another (nameless) image annotation system based on ontologies, making it possible to apply machine logic to the dataset using description logic (Hu *et al.*, 2003)[24].

2.1.5 Adding annotation-as-knowledge: ontologies

This section describes ontologies to define the meaning of annotation-as-thing.

A definition of ontologies is given by Mädche and Staab (2001)[25]: "Ontologies are (meta)data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine processable semantics. By defining shared and common domain theories, ontologies help both people and machines to communicate concisely, supporting the exchange of semantics and not only syntax."

Besides ontologies, many terms are used to structure concepts and relationships, such as lexicons, controlled vocabularies, thesauruses and taxonomies. They differ in how much meaning is specified to each term, and the language that is used to specify that meaning (Pidcock, 2003)[26]. Ontologies provide the richest description and are provided in a machine-processable language; hence, computer-reasoning ("if a disease occurs on fingers, it occurs on hands as well because fingers belong to the hand") is only possible using ontologies.

To make this desired application more intuitive, consider the example in Figure 8, which is a very simple ontology of the anatomical structure of the hand. When working with ontologies, a distinction needs to be made between <u>classes</u> (*finger* and *hand* are both classes; *finger* is a subclass of *hand*) and <u>instances</u> of classes (*left hand* and *right hand* are instances of the class *hand*).

[g] Not the device build by Apple

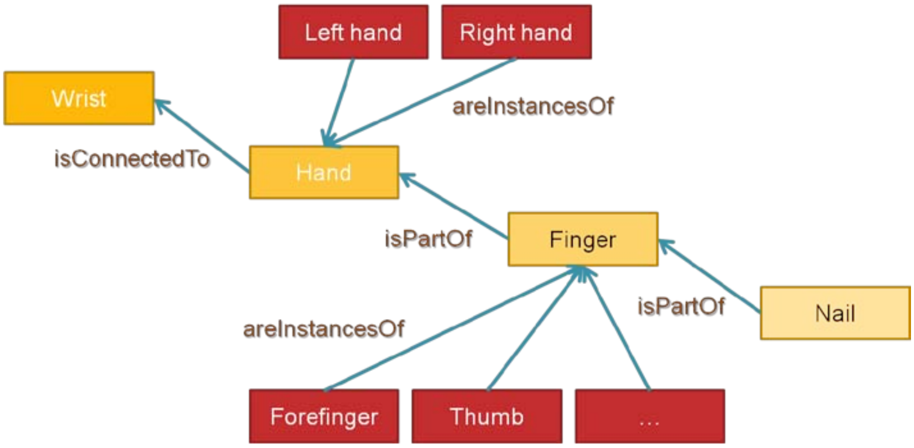


Figure 8 - A very simple anatomical ontology of the hand

Biomedical science is on the forefront in using and researching ontologies, but these ontologies are far from completely incorporated into their IT systems. Some of the available ontologies are general ones such as WordNet, more specific ontologies such as the Foundational Model of Anatomy, or very domain-specific ones as ONTODerm, RadLex and Gene Ontology. Ontologies are also used to guide data in the Electronic Health Record, for instance using SNOMED. A less advanced code system for EHR's is ICD, but this is a code system (taxonomy) rather than an ontology. All these systems will be discussed in detail further on in this chapter.

2.1.6 Standards for information exchange and imaging

Various IT systems need to be able to communicate with each other, therefore agreements on how such communication happens needs to be made. Such agreements are called standards. Several of such standards have been developed, of which some will be discussed throughout this chapter.

2.2 Technical standards for semantic information exchange

Several standards have been developed and are being used widely to define knowledge, order that knowledge hierarchically and to reason with that knowledge. This section describes the building blocks needed to store and transmit semantic information.

2.2.1 W3C and the Semantic Web

The World Wide Web Consortium (W3C)[h] is an international community that develops open standards to ensure the long-term growth of the Web[27]; one of their research topics is the Semantic Web. Using their standards for the Semantic Web, semantic information exchange can be accomplished. Their standards are widely used.

2.2.2 XML: Extensible Markup Language

When looking at ways to store and transmit data, the Extensible Markup Language (XML) cannot be ignored. This standard is developed by the W3C and is used to display structured data in plain text. The XML homepage provides the following example:

```xml
<?xml version="1.0"?>
<note>
 <to>Tove</to>
 <from>Jani</from>
 <heading>Reminder</heading>
 <body>Don't forget me this weekend!</body>
</note>
```

The text describes how a note could look like in XML. This particular note is written by *Jani* (between the `<from>` tags), directed to *Tove* (`<to>`), titled *Reminder* (`<heading>`) and contains the text *Don't forget me this weekend!* (`<body>`). It is entitled *Note*.

The ability to define data structures in XML resulted in hundreds of XML-based languages[28] including RSS and XHTML. Several languages related to ontologies have been developed using XML, including DAML, OIL and OWL[29]. XML is commonly used to describe image annotations.

2.2.3 RDF: Resource Description Framework

The Resource Description Framework (RDF)[30] is another programming language developed by the W3C. Where XML is designed to give full flexibility to programmers, RDF is designed to make a structure that is machine-interpretable.

2.2.4 DL: Description Logic

Description logic (DL) is a family of formal knowledge representation languages. It is mainly used when providing logical formalisms, such as when constructing ontologies. Such logical formalisms in the form of ontologies are used a lot when coding knowledge, including in the medical world.

[h] www.w3.org

A machine is able to reason with knowledge expressed in DL. The following medical example is derived from Schulz *et al.* (2010)[31]. The disease Hepatitis can be defined as 'the class of all instances that belong to the class *Inflammatory_Disease* and are further related through the relation *has_location* to some instance of the class *Liver*'. In DL notation, this statement would be expressed using

$$Hepatitis \equiv Inflammatory\_Disease \sqcap \exists has\_location.Liver$$

2.2.5 OWL: Web Ontology Language

The Web Ontology Language OWL is designed for use by applications that need to process the content of information, instead of just presenting information to humans[32]. It is a revision of the DAML+OIL language[33]. It is designed by W3C and is build upon XML and RDF.

For the Semantic Web, the first level above RDF is an ontology language that can formally describe the meaning of terminology used in data. When machines are expected to perform useful reasoning task to this data, the language needs more resources to describe relationships than RDF. OWL has got these resources and is able to define relationships between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, enumerated classes and more.

OWL uses DL to perform reasoning. It has three modes (OWL Lite, OWL DL and OWL Full) to give users the flexibility to choose between ease of use, expressiveness, completeness and decidability.

Practically every large ontology is available in OWL and practically every new ontology is made in it.

Protégé-2000

Protégé-2000[34,35] is a free, open-source platform that is designed to view, build and edit ontologies. It is initiated and licensed by the *Stanford Center for Biomedical Informatics Research*. It can import ontologies written in OWL and can export in XML, RDF and OWL. Listing over 200,000 registered users (July 2012), it is by far the largest tool for ontologies.

2.3 Standards for data exchange and imaging

Several standards have been developed to ensure interoperability in the medical domain.

2.3.1 HL7: Health Level 7

Health Level 7 (HL7) is an international standard for the electronic exchange of medical, administrative and financial data between healthcare IT systems. It is developed by the eponymous organisation. In Dutch medical domain, it is the only standard of which there is complete alignment to.

Version 2 of the system (HL7v2) is being used in practically every hospital in the Netherlands. Its latest version (HL7v3) is used in most ongoing developments, such as the new national infrastructure for the exchange of patient data called Aorta[36] by the Nictiz.

2.3.2 DICOM

Where HL7 is a general standards for digital information exchange in the medical world, DICOM is designed specifically for imaging.

DICOM stands for Digital Imaging and Communications in Medicine and describes how medical images should be stored, exchanged and printed. It is developed by the American National Electrical Manufacturers Association (NEMA)[i].

The following example describes (only a fractions of) the advantages of DICOM is over existing imaging standards such as JPEG. Think of an MRI image[j] that is being sent from the physician that is operating the MRI machine to the treating doctor. Such an image is constructed out of dozens images of slices of the body; these are all contained in one DICOM file. The physician that made the image can also add comments to the file, such as if the person was dizzy or restless in the MRI machine.

DICOM is used in practically every domain where imaging plays a serious role. An exception though seems to be dermatology.

Challenges for DICOM

Gibaud (2001)[5] examined the current status of standards for biomedical imaging. He states that DICOM is not widely used in research labs because it does not support XML for annotation. He concludes that "[...] the sharing of images and related data [...] remains an open question. [...] Substantial work remains to be done to re-express image-related information using ontologies."

Dermatology in DICOM

Also for dermatological imaging there are open questions. Little over a decade ago, DICOM Working Group 19 was created to address these needs of dermatologic imaging. It had to cancel its activities prematurely because "it was not possible to engage the relevant vendors, professional societies, and other interest groups in the effort." (Madden, 2011; unpublished work)[16]. Madden argues that times have changed and new efforts should be taken to develop a dermatological standard for DICOM. For the time being, no standards for dermatologic imaging exists.

One other reason why DICOM is not integrated into the dermatological domain is that a lot of photographic images are made using simple consumer cameras; hence there is little reason to use DICOM. However, upcoming techniques such

---

[i] www.nema.org

[j] A great set of example images can be found at www.osirix-viewer.com/datasets/

as digital dermascopy, body mapping and polarisation, wavelength filtering would benefit from a dermatological standard in DICOM.

2.3.3 IHE: Integrating the Healthcare Enterprise

The Integrating the Healthcare Enterprise (IHE) initiative began in 1998 as an effort to define more clearly how existing standards, especially DICOM and HL7, should be used to resolve common tasks for information system communication. The IHE technical framework defines, precisely, a common information model and a common vocabulary for systems to use in communicating medical information. It then specifies, precisely, how DICOM and HL7 are to be used by information systems to complete a set of well-defined transactions that accomplish a particular task. At the same time, the framework provides a common human vocabulary that professionals and vendors can use to discuss further problems of this nature (Siegel and Channin, 2001)[37]. Twenty-one hospitals and about seventy companies and authorities are involved in the Dutch IHE[k], including the LUMC, the developers of eZIS (Chipsoft) and CA (RVC), the Dutch HL7 and the Nictiz.

[k] www.ihe-nl.org

2.4 Standard terminologies, code systems and ontologies

Many standard terminologies exist in the medical domain and many code systems have been developed for them. This section describes systems that are used in EHR's.

Every system is developed for a different purpose but there is overlap between them. To name one, cross mappings exist between ICD and SNOMED, for instance: in ICD-10 the terms for *inflammation* are classified as *inflammatory diseases*, whereas SNOMED III has *inflammation* under a separate taxonomy containing properties or structures produced by an *inflammatory disease*[38].

2.4.1 International Classification of Diseases

The International Statistical Classification of Diseases and Related Health Problems (ICD) is an international list of diseases. It is developed by the World Health Organisation (WHO)[l]. It is one of the two major classification systems for EHR's and used by the LUMC in eZIS.

The first version was introduced in 1900. The mostly used version of ICD in the Dutch healthcare system is ICD-9, released in 1975. The latest version is ICD-10 (1994), but because it contains many expansions in comparison to the previous version, is not possible to convert version 9 directly into version 10; as a result there are a lot of hospitals unwilling or unable to upgrade to version 10. The World Health Organisation is currently revising the ICD towards the eleventh version; its release is expected in 2015. To improve interoperability between systems, ICD-11 will contain direct links to ontologies such as SNOMED CT.

Several variations of ICD-9 are developed, such as ICD-9 DE (Dutch Extension) and ICD-9 CM (Clinical Modification, providing additional morbidity detail). The latter is used by the LUMC.

Dermatology in ICD

Little dermatologists were involved in the development of ICD-9, resulting in a low number of available classes to describe dermatological disorders. The LUMC dermatology department made an extension to ICD-9 for internal use, adding extra subclasses. For instance, the class *Malignant dermatoses* is divided into eleven subclasses, including *lymphoma* and *Paget's disease*.

The International League of Dermatologic Societies (ILDS)[m] states on its website that also in version 10 of ICD "distinct diagnoses are quite often classified under a rather crude denominator. In addition, many important terms used by dermatologists are completely missing. As a consequence ICD-10 codes do not sufficiently fulfil the fundamental requirements for documentation of dermatological diagnoses for statistical and scientific purposes."[39]

2.4.2 SNOMED CT

The other major classification system used in EHR's is the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT, or SNOMED in short). It is a collection of medical terms for medical codes and synonyms. Using ontologies to form its structure, the information is computer processable[40]. It contains over 250,000 classes named by over 400,000 terms. Containing a wide collection of medical terms and being computer processable, it is a much more sophisticated system than ICD is.

---

[l] www.who.int/en/

[m] web.ilds.org

SNOMED is developed by the International Health Terminology Standards Development Organisation[n] (IHTSDO). In January 2002 SNOMED CT was created by the merger of SNOMED RT (Reference Terminology) and Clinical Terms version 3. SNOMED CT combines the historical strength of both: SNOMED was well known for its coverage of medical specialties, whereas the strength of Clinical Terms version 3 were its general practice terminologies. SNOMED CT is currently being used as the ontological basis of the upcoming ICD 11.

Nictiz describes SNOMED CT as 'means to come to unity of language' and 'the first step towards unambiguous information[41]' and promotes its use throughout the Dutch healthcare system.

Dermatology in SNOMED CT

Also in SNOMED CT, the possibilities to describe dermatological disorders are insufficient. The DermLex project (described in the next section) aims to develop a code system that can eventually be incorporated into SNOMED.

2.4.3 Dermatology terminology

Like every medical domain, dermatology uses a specific jargon (semantics). Data structures can be adapted to this terminology resulting in improved information retrieval and exchange.

Efflorescences

The information in this paragraph is derived from the handbook for dermatology used by LUMC students that are doing their residency at the dermatology department[42].

A good dermatologist is a good describer of skin diseases: accurately describing affected skin areas results in better patient treatment and better research material. The terminology to describe this morphology is by naming the efflorescences: changes in the colour, appearance and texture of the skin. One can distinguish between primary and secondary efflorescences: primary efflorescences are the changes that come directly from the pathological process in the skin, secondary come from external processes, such as scratching or an infection. Primary efflorescences are divided into twenty-three distinctive classes, including macula, papules, nodules, pustules and plaques.

Other important features of the disorder to look after are place on the body, arrangement (number and distribution of spots), size, shape and outline of individual spots and changes in skin complexion and skin surface.

The PROVOKE nomenclature

Van Everdingen[43] designed a nomenclature that guides the process of describing skin disorders, called PROVOKE. It is an acronym for the Dutch letters used to describe the features stated in the previous paragraph: Plaats (place), Rangschikking (arrangement), Omvang (size), Vorm (shape), Omtrek (outline), Kleur (colour) and Efflorescenties (efflorescences). In practice only few dermatologists use it in such an extensive way, still computer systems can take advantage of the nomenclature.

PROVOKE only used in Dutch-speaking hospitals.

DermLex: Dermatology Lexicon Project

DermLex[44] is a dermatology lexicon developed by the American Academy of Dermatology (AAD)[o]. Version 1 was released only in 2009. The most important goal of DermLex is to create a common language among dermatologists and

[n] www.ihtsdo.org

[o] *www.aad.org/dermlex/History.aspx*

between specialties. Eventually, DermLex should have online tools so it seamlessly integrates into EHR's such as SNOMED CT[45].

ONTODerm

In 2008 an initial attempt has been made to build ONTO*Derm*[p]: a domain ontology for dermatology. The creator states that the system is "a step towards semantic dermatology. [...] It is by no means a complete and stable ontology [but] an ongoing project." Physicians and software developers are invited to participate in the development.

ONTO*Derm* concentrates on six important concepts for which dermatology is different from other medical specialties, such as the description of lesions and patients, and investigation techniques. The goal of ONTO*Derm* is to integrate into various other software systems. (Eapen, 2008)[46].

2.4.4 Unified Medical Language System

The Unified Medical Language System is a repository of biomedical vocabularies developed by the US National Library of Medicine[q]. It integrates over 2 million names for some 900,000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. It tries to overcome two significant barriers to effective retrieval of machine-readable information: the variety of names used to express the same concept and the absence of a standard format for distributing terminologies. It integrates many sub domains including SNOMED, OMIM and GO (both related to genes), MeSH (related to literature), UWDA (anatomy) and NCBI (organisms).

Since the UMLS is a repository rather than a terminology it is not used in classification systems such as EHR's. One possible application is using the UMLS as a controlled dictionary, such as done by Woods *et al.* (2004)[47] or Tolentino *et al.* (2007)[48].

2.4.5 Reference ontologies

The following section describes ontologies that are designed to serve as reference systems: they might not be directly applicable to the biomedical domain but can contribute to the success of other systems.

Foundational Model of Anatomy

The Foundational Model of Anatomy (FMA) is an ontology focussing exclusively on the representation of structure; as a result it can serve as a reference ontology for other ontologies of which anatomy is a component. The initial development is an enhancement of the anatomical content of the UMLS[49,50].

It contains nearly 70,000 concepts and is implemented in Protégé. Definitions of the FMA have been used as a basis for characterising definitions of anatomical concepts in WordNet[51] and for other ontologies, including the UMLS[52].

WordNet

WordNet is a lexical database that serves as a resource for applications in natural language processing and information retrieval. It contains most English nouns, adjectives, verbs and adverbs. WordNet is based on a set of synonyms (*synset*) that represent one underlying concept. This information is based on synonymy (one meaning expressed by several words) and polysemy (one word having several distinct meanings). It is developed at Princeton University. The current version of WordNet (2.0) contains over 114,000 noun synsets categorised into nine hierarchies[50,53].

---

[p] *ww.gulfdoctor.net/ontoderm/*

[q] *www.nlm.nih.gov*

The success of WordNet is largely due to its accessibility, quality and potential in terms of natural language processing. Some applications are the structuring and categorisation of documents, and improvements in natural language processing systems[54].

Wikipedia[55] provides an example of information that is stored in WordNet: "The word *dog* would have the following hypernym hierarchy; the words at the same level are synonyms of each other [...]:

dog, domestic dog, Canis familiaris
- ➔ *canine, canid*
  - ➔ *carnivore*
    - ➔ *placental, placental mammal, eutherian, eutherian mammal*
      - ➔ *mammal*
        - ➔ *vertebrate, craniate*
          - ➔ *chordate*
            - ➔ *animal, animate being, beast, brute, creature, fauna*
              - ➔ *..."*

PART 2
OTHER INFORMATICS TECHNIQUES

Part 1 described techniques to achieve semantic interoperability. Often semantic interoperability needs other techniques in order to result in benefits for healthcare; this section covers Information Extraction, Data Mining, Machine Learning and Computer Vision.

> **Other techniques from computer science**
>
> ❖ Machine Learning (using Computer Vision)
>
> ❖ Information Extraction
>
> ❖ Data Mining

2.5 Machine Learning & Computer Vision

This section describes Machine Learning using Computer Vision (ML/CV), that can be used to have a computer annotate images automatically. Consequently, ML/CV can be used for computer-aided diagnosis and to decrease the effort of manual image annotation.

Machine learning (ML) allows machines (computers) to 'learn' from datasets. The simplest example is that after giving a computer the sentence "1, 2, 3", the computer knows that the next item in the sequence will be "4". Where Data Mining tries to find previously *unknown* information, Machine Learning aims at prediction based on *known* data. Computer Vision (CV) is the field that concerns with understanding and analysing images. Together, they are frequently used to support the automatic annotation of images (for example Duygulu *et al.*, 2002[56]; Jeon, Lavrenko and Manmatha, 2003[57]). Attempts have also been made in the medical field, although such systems are mainly built to classify X-Ray images (for instance Yao *et al.*, 2006[58]; Kalpathy-Cramer and Hersch[59], 2007; Mueen *et al.*, 2008[60]). A review of systems that are made for the medical domain is written by Müller *et al.* (2004)[61].

Such approaches are usually designed to classify high-level semantic features (such as a house or person in the general case), or to determine the modality of an image (whether it is a X-Ray image, CT scan, microscopy photo, etc.); therefore challenges lie in the detection of more low-level features such as the body part or skin complexion.

2.5.1 Requirements for proper ML/CV

Machine learning poses some requirements to images. Haas (2012, unpublished work[62]) examined the possibilities at the LUMC dermatology department to classify dermatological images using ML/CV. The following conditions to improve the quality of Machine Learning algorithms in the department are suggested:

- ❖ Information about the level of privacy of each image should be stored;
- ❖ A plain background should be used when taking images;
- ❖ Information about the size of the object in an image should be annotated;
- ❖ Ditto for the skin complexion of the patient.

2.5.2 Possibilities

This section explores several possibilities for automatic image annotation at a dermatology department, using ML/CV. Dermatological images are almost entirely images with a micro view (containing only one person or even only one body part, in front of a plain background), these are perfect conditions for Computer Vision techniques.

Modality detection

There are only two relevant modalities in dermatology: a regular or a dermatoscope photograph. No research on the detection of these different modalities area could be found, but the fact that computers are able to distinguish two very similar modalities (X-Ray and CT) suggests that distinguishing between images made by a dermatoscope and a regular photo camera is feasible.

Skin complexion detection & classification

Research done on skin complexion detection is mostly focussed on face detection, face recognition, and content filtering[63]. Hence, most techniques are aimed *detecting* skins, whereas dermatology is in need of *classifying* skin complexion into categories in for instance SNOMED. No research could be found on the classification of skin complexion.

The *detection* of skin is very important for the success of other techniques, as will be described in the next section. Real-time detection of skins has been possible since 1999 [64,65,66,67]. The amount of true positives in skin complexion detection can reach up to 94.7% with false positives 30.2%[68]; others proved it possible to accomplish 80% true positives with only 8% false positives[69].

Lesion border detection

Another critical step in applying ML/CV is the detection of lesion borders[70]. Most techniques focus on the lesion border detection in dermascopy images (for instance Celebi *et al.*[71]), others on regular images (for instance Haas[62]).

Computer-aided detection and diagnosis

Computer-aided diagnosis (CAD) are procedures in medicine that assist doctors in the interpretation of medical images[72]. Dermatology is beginning to research these techniques, especially on melanoma and pigmented skin lesions[73]. Most CAD techniques focus solely on dermatoscope images.

Already in 1993, the first steps were taken on the computer-aided discrimination of malignant melanoma[74]. It has been possible for a computer to discriminate 95% of the images correctly (2011)[75]. Advancements are also made in other areas, such as the automatic detection of blue-white veil (2008)[76].

Haas (2012, unpublished work)[62] showed that it is possible to discriminate between four visually dissimilar disorders, suggesting that CAD may be possible for more disorders.

Limb recognition

The classification of human body parts – mainly limbs – seems to be an inactive field of research. In the late 1990s some attempts have been made to recognise humans, especially for surveillance systems, but most systems focus on detecting human action (motion) or posture (for instance Ghost[77], Ghost3D[78] and "W^4" [79], all by Haritaoglu et al (1998, 1998, 2002, respectively); or the nameless systems by Polana *et al.* (1994)[80] and Overney *et al.* (2005)[81]).

Most of these systems try to recognise humans in images with a macro view: images that can contain multiple persons as well as a scenery. As most dermatological images are images with a micro view, Template Matching and Keypoint Matching techniques (both part of Computer Vision) can be useful to detect limbs. No literature could be found on limb recognition using these techniques, so future research is needed.

2.5.3 The CLEF Automatic Medical Image Annotation Task

A famous test to review the quality of automatic medical image annotations is the annual CLEF Automatic Medical Image Annotation Task (for instance 2005[82], 2009[83]). Because this test consists of almost solely radiography images, it is of no use to us.

2.6 *Text Mining in Electronic Health Records*

Text Mining is a very powerful tool that is used to extract knowledge from unstructured data, as defined by Hearst (1999)[84]. Typical aspects of Text Mining are Information Extraction (to extract specific of information from texts) and Data Mining (to find associations among the extracted pieces of information).

2.6.1 Information Extraction

Information Extraction (IE) can be used to extract known facts from data. Often such data are input using free-text fields: text fields in which a user is able to freely type, such as a 'comment' field'.

DeJong (1982) states that IE can only extract "predefined types of information from text"[85], not any information. Such predefined types of information can be for instance lexicons or templates. Meystre *et al.* (2008)[86] state that extracting information from clinical texts is a challenge because many texts are telegraphic, contain many shorthand and misspelling, and long lists of laboratory values or vital signs can be a major distraction.

When IE techniques are able to perfectly extract information from large texts, this takes away the necessity for data input to occur via many different fields, resulting in novel user interfaces.

Independent from changing user interfaces: the retrieval of information can benefit greatly from IE. For example, when a physician has entered some PROVOKE features such as the place and outline of a wound into a free text field, an IE algorithm (that 'understands' PROVOKE) might be able to notice that this place and outline is described, hence enabling the ability to search for them.

Meystre *et al.* reviewed the current status of IE research on clinical text and conclude that "results in clinical text IE [are] often mixed. [...] More experience is needed, annotated text corpora are rare and small [because of privacy issues], and clinical text is simply [hard] to analyse. [...] During the last several years, performance has gradually improved, exceeding 90% sensitivity and specificity in several cases. Systems are now mostly statistically-based, and therefore require annotated corpora for training. Creating annotated clinical text corpora is one of the main challenges for the future of this field".

ICD, SNOMED CT and the UMLS have been used to support Information Extraction for instance in the "2007 international challenge: classifying clinical free text using natural language processing" [87].

Other possibilities lie in the field of Information incompleteness and Information integration, as summarised by Ramakrishnan, Hanauer and Keller[88]:

Information incompleteness

In order for *Data Mining* (see the next section) to take place, the data first needs to be complete. Data Mining might not detect the association between asthma exacerbation and smoking when asthma exacerbation is coded but smoking is only described in clinical documents.

Information integration

Information extraction is able to connect information such as laboratory finding, genomic tests, numerical and other physiological data – information that is currently stored in different systems.

2.6.2 Data Mining

Data Mining techniques can be used parallel to those for Information Extraction. In contrast to IE in which known facts are extracted from data, Data Mining is involved in finding *any* useful information in data. Data mining requires a training set.

Data mining techniques are interesting in relation to EHR's. Ramakrishnan, Hanauer and Keller (ibidem) review many of its potentials; two of them are sequential modelling, and verification and validation.

Sequential modelling

Much of the patient's interaction with a hospital is temporal, but symptoms and diagnoses always develop sequentially. Hence, the mining of sequential patterns may serve as the gathering of information to tune a hospital's infrastructure.

Verification and validation

Comparing the outcomes of a Data Mining algorithm on two different datasets, for instance of two medical institutions, might find interesting differences.

PART 3
PRELIMINARY CONCLUSIONS

Now that all systems and standards have been discussed, the subquestions related to the current situation at the department and of the techniques can be answered (questions 1-3). In the next chapter question 4 about the recommended future situation of the department will be answered.

1. The current situation at the LUMC dermatology department

This section discusses the current situation at the dermatology department of the LUMC.

1.a. What information systems are being used, especially related to handling images?

The Hospital Information System that is used is called eZIS. This system contains the Electronic Health Record and Diagnose-Behandelcombinatie. The PACS is called Clinical Assistant. The code system used to describe diseases is ICD-9 CM with a modification related to dermatology.

1.b. How is the accomplishment of semantic interoperability being handled?

The PROVOKE nomenclature (including efflorescences) is used to describe dermatological disorders; the DICOM standard is not used as it is contains no dermatological guidelines. The LUMC relies on Health Level 7 for its digital communication. Images are hardly directly annotated; when they are the annotations are drawn directly to the images. Indirect annotation is insufficient as many information is unconnected.

1.c. How do the staff judge the usability of these systems and techniques?

ICD-9 CM is insufficient and is modified for internal use. The HIS is found inconvenient as it disrupts the workflow of medical staff when examining patients, relevant information is not shown and stored information is hard to retrieve. The dermatology department is financially disadvantaged by the DBC system; this causes dissatisfaction.

2. Techniques to achieve semantic interoperability

This paragraph discusses the basic techniques used to integrate semantic knowledge.

2.a. What is semantic interoperability?

The term interoperability explicitly mentions the ability to reuse data. It is one of the key aspects of computer science as it is a requirement for many informatics technique to succeed. Semantic interoperability is reusing data such that the meaning of it is unambiguous and shared.

2.b. What are the basic techniques to accomplish semantic interoperability?

To achieve semantic interoperability, data need to have a meaning to a computer. This can be accomplished by using a decent computer-processable nomenclature, annotated data, and standards for information exchange and imaging.

2.c. What standard terminologies and code systems are available?

The standard terminology used to describe dermatological disorders in Dutch hospitals is the PROVOKE nomenclature.

The two main code systems used to describe disorders are ICD (version 9 and 10) and SNOMED CT; the latter is based on ontologies which allows computer-reasoning with data expressed using SNOMED. Dermatological terms are insufficiently embedded in any code system. ONTO*Derm* and DermLex are initiatives to capture dermatology terminology and implement it in code systems.

RDF, OWL, Protégé and Description Logic paved the way for ontologies and its key-feature: machine processibility.

The UMLS is a repository of many medical description systems; FMA describes the anatomical structure of the human body. WordNet contains most English nouns, adjectives, verbs and adverbs and is based on a set of synonyms.

2.d. What tools and techniques are available for data annotation?

Annotated information is made possible by XML, a computer-processable terminology system such as SNOMED CT, as well as knowledge about the human anatomy using the FMA. Machine Learning and Computer Vision allow information to be annotated automatically. Playing a computer game can transfer the effort of manual annotation from physicians to for instance medical students. Graphical tools that support the annotation process of medical images are available.

One has to keep in mind that the annotation process is time-costly and the benefits of it may not directly be visible. The benefits of thoroughly annotated data are less for physicians than for researchers; this makes it even more difficult to convince physicians to annotate their data.

2.e. What standards for information exchange and imaging are available?

A standard that guides the digital information exchange in hospitals is present in the form of HL7.

DICOM is available for imaging. It is capable of cooperating with HL7, but support for annotation (in terms of XML) as well as for dermatological imaging need improvement. The Integrating the Healthcare Enterprise initiative promotes the use of both HL7 and DICOM and is responsible for a major part of the integration of both techniques into the medical domain.

3. Supportive techniques

3.a. What other techniques that support t the accomplishment of the desired benefits are available?

The four supportive techniques that are covered in this thesis are Computer Vision, Machine Learning, Information Extraction and Data Mining.

In combination with Computer Vision, Machine Learning is able to find patterns in data and to predict how other data fit into that pattern; thus when a training set is available it can automatically annotate images with a certain precision. Possible annotations are skin colour classification, limb recognition and lesion border detection.

Information Extraction can result in better user interfaces and can be used to find incomplete information and integrate such information. Data Mining can be used for sequential modelling and verification and validation of data.

3.b. How usable are these techniques?

It is possible to discriminate between some skin disorders, but much more research is needed in this field of research. A lot of research has been done on the detection of limbs in images with a macro view, but surprisingly little on images with a micro view. Template matching is a technique that has a high potential for success in detecting limbs in images with a micro view. Further research needs to be done on the automatic classification of skin complexions, into for instance classes in SNOMED.

Work on IE needs to continue as no stable, consequent systems are available yet. Hospital data sets need to be made suitable for Data Mining.

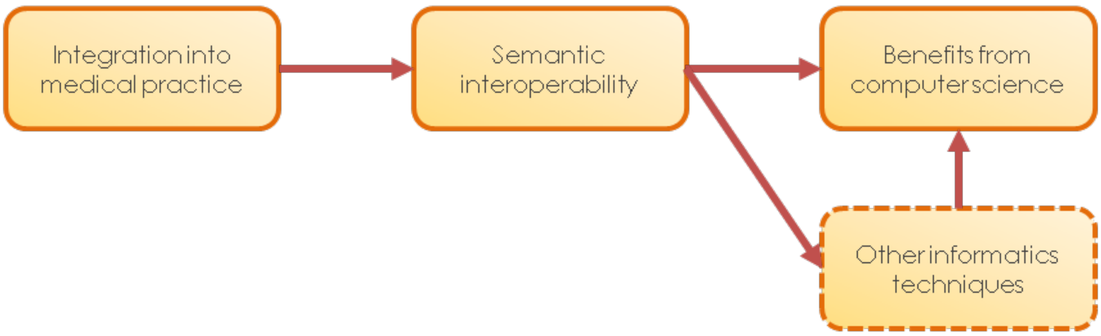
4. The future of the LUMC dermatology department

So far, all relevant systems and standards that are available have been covered. The next chapter will integrate all these into a recommendation for the department, thus answering the last two subquestions.

3. READY FOR THE FUTURE

Replacing a large IT system as a result of any paper is unlikely and some of the techniques described in the previous section have great potential but have not yet matured; therefore the goal of this article is to recommend what the department needs to do in order to ready for the future in health informatics. Hopefully will this thesis encourage people to reflect on their current systems and will motivate them to put more time into achieving semantic interoperability: the efforts will eventually be worth it.

This chapter covers how a proper basis can be designed that is ready for the future in health informatics. The structure suggested in chapter 1 will be followed:



3.1 Step 1 – Integrating the basic techniques into the medical practice

This section produces a list of recommendation to integrate semantic interoperability into the medical practice. This is a requirement for many other techniques and will eventually result in better health care, research and education.

The integration into the medical practice is a difficult tasks and requires a close cooperation between computer scientists and the medical staff.

Most attention is paid to data annotation as it is likely to have the biggest impact on the workflow of medical staff.

Integration in the medical practice

- Switch to SNOMED CT
- Guidelines for image annotation
- Division of labour
- Automated annotation
- Supportive graphical interface
- Keep on using HL7 and orientate on DICOM
- Graphical interface to connect CA and eZIS

3.1.1 Decent computer-processable nomenclature

A standard terminology decreases the risk of error by miscommunication. The usage of a code system is inevitable when working with digital systems such as the Electronic Health Record; a good coupling between this terminology and the code system therefore is important.

A terminology is present in a system to describe efflorescences (in the Netherlands embodied in the PROVOKE nomenclature).

There are two reasonable code systems available on the market: SNOMED CT and ICD. SNOMED CT is the most comprehensive and sophisticated medical nomenclature that is available; being based on ontologies, it makes it possible for machines to reason with any data expressed in it. SNOMED has the highest potential of becoming the lead system in the medical world. The National IT Institute for Healthcare in the Netherlands promotes the use of SNOMED.

The PROVOKE nomenclature is insufficiently integrated in any code system. The development of DermLex and ONTO*Derm* is very important for the semantic knowledge integration in the dermatological domain thus the entire sector should contribute to a quick complement and integration of these systems.

Making a cost-benefit analysis (CBA) of the usage of each system is a very complex process and such a CBA will always contain assumptions and uncertainties, therefore it is out of scope of this article. The terminology system that is currently used (ICD-9) dates from 1977 and an upgrade to any new system will be needed in the near future. Because cross-mappings exist between ICD-9 and SNOMED, the costs of upgrading to any of these two systems will be comparable. The benefits of relying on SNOMED CT instead of ICD will be numerous for patient care, research and education, and eventually outweigh the effort and costs of switching to a new system. Therefore, the LUMC is advised to switch to SNOMED CT.

While ICD-9 is used, effort should be put in fully integrating it into every IT system, including the DBC. While ICD-9 and CA are still used, the department can contact the makers of the applications to talk about less-rigorous steps than switching to a new code system.

- *Recommendation 1:* *The LUMC should fully integrate SNOMED CT into its digital systems.*
- *Recommendation 2:* *While ICD-9 is used, effort should be put in fully integrating it into every IT system, including the DBC.*
- *Recommendation 3:* *While ICD-9 and CA are still used, the department can contact the makers of the applications to talk about less-rigorous steps than switching to a new code system.*

3.1.2 Data annotation

Data annotation is the second important point in integrating semantic knowledge.

Researchers demand annotations of a higher quality than physicians do; to prevent unnecessary work the purpose of one annotation should be kept in mind.

- *Recommendation 4:* *Researchers demand annotations of a higher quality than physicians do; to prevent unnecessary work the purpose of one annotation should be kept in mind.*

Guidelines for annotation

Guidelines for annotation are a requirement for the integration of image annotation into the medical practice.

Not only should new image material be annotated shortly after the image is produced, effort should be put in annotating old material as well. When possible, physical annotation should be replaced by graphical annotation. The usage of an old nomenclature like ICD-9, or the fact that data in the DBC are converted to a less advanced code, should be no excuse not to annotate

- *Recommendation 5:* *New image material should be annotated shortly after production. The usage of an old nomenclature like ICD-9, or the fact that data in the DBC are converted to a less advanced code, should be no excuse not to annotate*
- *Recommendation 6:* *Effort should be put in the annotation of old material.*
- *Recommendation 7:* *Physical annotation should be replaced by graphical annotation when possible.*

An image annotation should be in XML format. If possible it should map to classes in for instance SNOMED CT or the FMA. As many information as possible should be annotated, such as age, sex, skin complexion and living place of the patient, diagnose, the primary body part, orientation and size of the photographed object, stage of the treatment and colour of the background. If the staff accidentally do not have enough time to annotate everything, at least the patient information, diagnose, skin complexion and primary body part should be annotated. It should be kept in mind that non-doctors might need to work with the images, stretching the need to annotate whether the image should be classified for non-doctors.

The annotation of the maximum amount of available information allows the doctor of the example in chapter 1 to query all images that have an instance of *Spinocellulaircarcinoom* that occurs on the hands, the public health investigator to examine whether changing legacies have health influence for people living in a certain area, and the computer scientist to train a classifier for Machine Learning.

The annotation of the following image could be[r]:

```
<?xml version="1.0"?>
<annotation>
  <primary_limb>
    <name>Hand</name>
    <fma_class>288524001</fma_class>
    <amount>2</amount>
    <orientation>palm_up</orientation>
  </primary_limb>
  <background_colour>light-green</… >
  <patient_info>
    <age>61</age>
    <sex>male</sex>
    <living_place>Leiden</living_place>
    <skin_compl_snomed_class>37943185</…>
 </patient_info>
 <disorder>
    <diagnose>unknown</diagnose>
    <treatment_stage>start</…>
  </disorder>
 <classified>false</classified>
</annotation>
```

Figure 9
An example image and its annotation

- *Recommendation 8:* *Image annotations should be in XML format.*
- *Recommendation 9:* *If possible, image annotation should map to ontologies like SNOMED CT and the FMA.*
- *Recommendation 10:* *As many information as possible should be annotated, such as age, sex, skin complexion and living place of the patient, diagnose, the primary body part, orientation of the body part, size of the photographed object, stage of the treatment and colour of the background.*
- *Recommendation 11:* *It should be kept in mind that non-doctors might need to work with the images, stretching the need to annotate whether the image should be classified for non-doctors.*

[r] The numbers are not corresponding to SNOMED and the FMA.

- *Recommendation 12:* *If the staff accidentally do not have enough time to annotate everything, at least the patient information, diagnose, skin complexion and primary body part should be annotated.*

Division of labour

Tools have been made to enable image annotation in the medical world, such as iPad. Its authors state that "iPad is targeted primarily for enabling imaging research [...]. In clinical practice, few cases require detailed image annotation [...]. [...] Enhancements to the user interface would be needed to make iPad practical to use in high-volume image interpretation paradigms, such as incorporating voice recognition [...]. (Rubin *et al.*, 2008)[23]" This suggest that it might not be a feasible goal to let physicians annotate images as part of their routine workflow and other solutions have to be studied.

The time of a physician is arguably more valuable than the time of for instance a hospital photographer; therefore a non-doctor such as a photographer should do the largest part of the annotation effort. When that person its medical knowledge is insufficient, he or she can consult a physician. The department is suggested to provide the person who annotates with extra medical education to minimise the amount of question he or she has to ask to the physician. This person is suggested to schedule the annotation-as-process at least weakly.

The division of labour should be very clear to prevent that there will not be annotated at all. It might be a good idea to let a physician supervise, or afterwards check, the first sets of annotations.

- *Recommendation 13:* *A non-doctor such as a photographer should do the largest part of the annotation effort. The division of labour should be very clear. When that person its medical knowledge is insufficient, he or she can consult a physician.*
- *Recommendation 14:* *The person who annotates should be provided additional medical educated. This person is suggested to schedule the annotation-as-process at least weakly.*

Only little research on how to minimise the annotation effort was found. Von Ahn and Dabbish (2004)[89] describe how images on the World Wide Web are annotated by individuals who are playing a game; the game-aspect makes that the players do not consider the annotation process as an effort. As medical images are confidential, letting 'the crowd' do the annotations is a bad idea. A solution is to let medical students – who already signed an agreement about professional secrecy – annotate these images, as part of their education. This lowers the annotation effort of medical staff and is likely to improve the annotation quality over annotations done by non-medically educated persons. Future research needs to clarify whether this is a good alternative to the manual annotation by a doctor.

- *Recommendation 15:* *Have medical students annotate images as part of their education, using a computer game.*

Automated annotation

Machines are not able to annotate everything and to annotate without making mistakes, but can do suggestions to support the process of manual annotation. An example is that a machine says "this photograph likely contains a hand (FMA class 288524001), and the skin complexion maps to SNOMED CT class 37943185". An interface that contains automated suggestions is explained in the next section.

It is likely possible for machines to classify skin complexion, classify the image modality, detect lesion border and recognize limbs within reasonable boundaries.

- *Recommendation 16:* *A classifier should be trained to give suggestions during the annotation process.*

Supportive graphical interface

For the success of integrating annotating into the medical practice, a graphical interface that supports the annotation process is essential as it lowers the effort of the annotation process. The interface can guide the annotation process and output the annotations in the desired XML format.

This section illustrates how manual annotating a body part on an image can look like when some of the techniques described earlier are being used. The image is a modification of a screenshot of Clinical Assistant. In this particular interface there are three options to annotate the primary body part, shown on the left part of Figure 10. This way the user can freely decide what option to choose. The supportive interface also allows non-doctors to annotate.

The three options to annotate the primary body part are:

a. By clicking on a region on the schematic representation of a body and then selecting the appropriate part from a list that is derived from the FMA (Figure *11*);
b. By typing the name of the body part and following the suggestions from the UMLS (*Figure 12*). The UMLS offer spell-checking support for many different terminologies, allowing the user to freely type text.
c. By following suggestions from a recommender that is trained using Machine Learning and Computer Vision (*Figure 13*).

- *Recommendation 17:* *Design a graphical interface that supports the annotation process. Allow the user to choose the input method of its choice.*

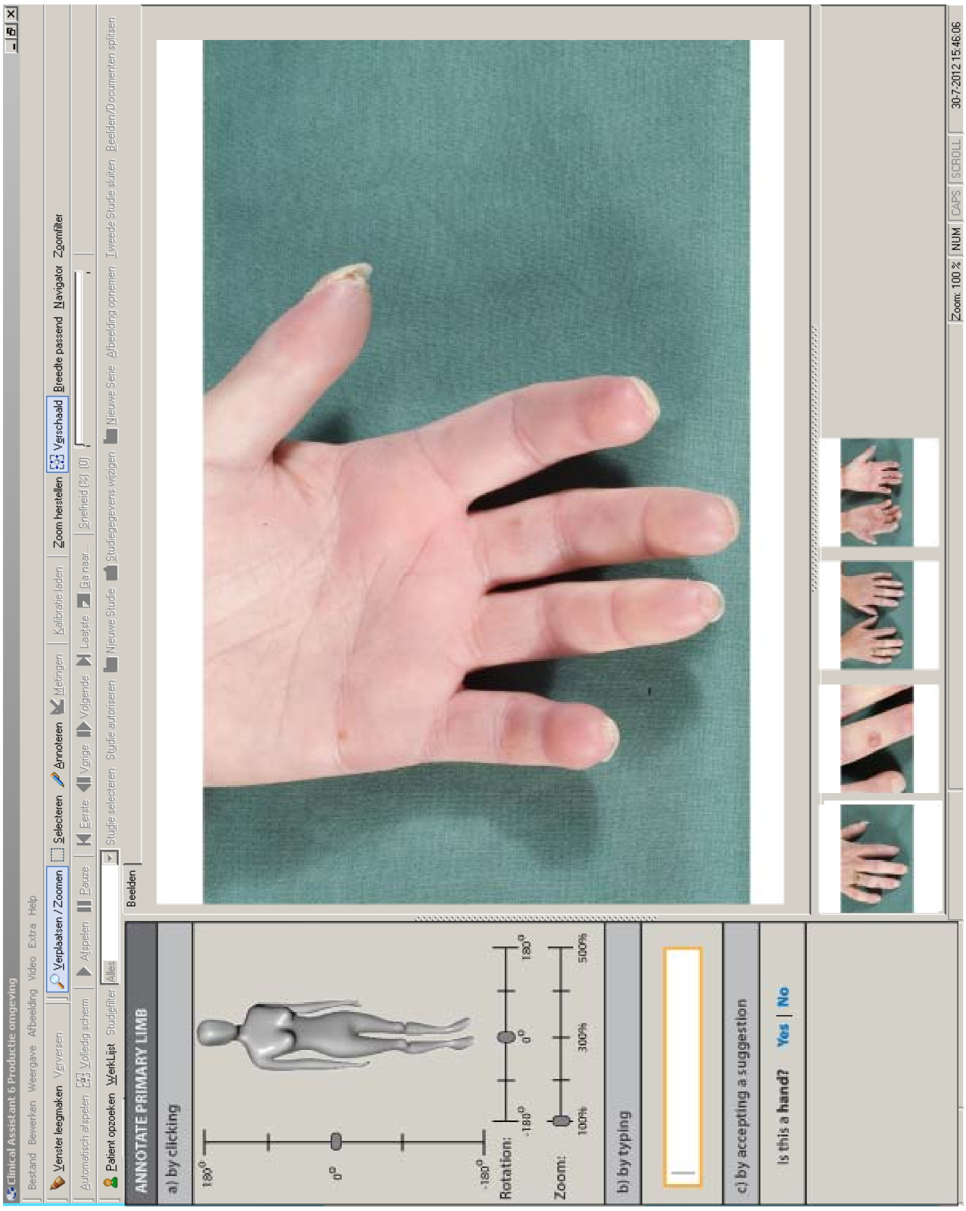
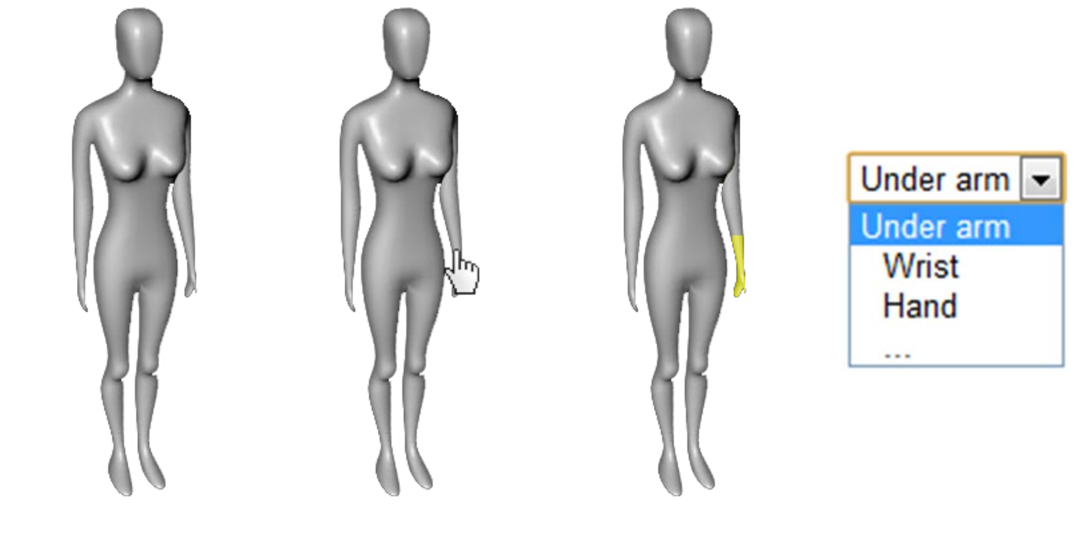

Clinical Assistant 6 Productie omgeving
ANNOTATE PRIMARY LIMB
a) by clicking
Rotation:
Zoom:
b) by typing
c) by accepting a suggestion
Is this a hand?
Yes | No

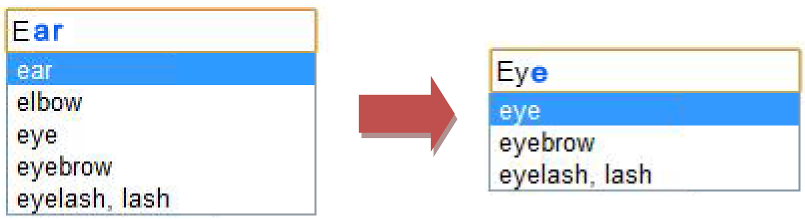
Figure 10



(i) *(ii)* *(iii)* *(iv)*

Figure 11 – Option a) of adding body parts:
By clicking on a place on the body and then selecting the appropriate part.

i) the clickable image of a human body; (ii) the user clicks on the left arm; (iii) the area of which the suggestions in (iv) are shown is marked yellow; and (iv) the program gives a list of suggestions.



(i) *(ii)*

Figure 12 – Option b) of adding body parts:
By typing the name of the body part and following the suggestions

(i) the user types in 'E' and gets a list of suggestions; (ii) the user types another letter 'y' after which the list of suggestions is updated. The user can keep on typing the name, or select one of the selections using the keyboard (arrows, tab and enter) and the mouse.

Figure 13 - Option c) of adding body parts:
By following the suggestion from the Machine Learning and Computer Vision algorithm

3.1.3 Standards for information exchange and imaging

The final step in integrating semantic knowledge is the use of standards. Health Level 7 prescribes how to transmit digital data; DICOM is a standard to describe medical images. Whereas HL7 is fully integrated in the medical domain, DICOM is not. Support in DICOM for annotation (in terms of XML) as well as for dermatological imaging need improvement

- *Recommendation 18:* *Keep on using HL7.*
- *Recommendation 19:* *Orient on DICOM and prepare the systems for it. As soon as dermatological standards come available for it, the LUMC should start using it. As long as no guidelines for image annotation are available in DICOM, recommendation 5-8 should be followed.*

Connection eZIS & CA

The example user interface of section 3.1.2 is designed to annotate the primary body part as part of CA. Other information, such as the orientation of the body part, and the lesion area, can also be annotated using this interface. Patient information, such as treatment history and patient data, are usually stored in the patient database eZIS. When standards such as XML for annotations, HL7 for information exchange and DICOM for imaging, are used, it is possible to design a graphical user interface that connects eZIS and CA, hence showing all relevant information in one interface (Figure 14).

- *Recommendation 20:* *Design a graphical interface over CA and eZIS to facilitate the process of manual image annotation. While this interface is not present, the department secretary should stay responsible for transferring information between the two systems.*

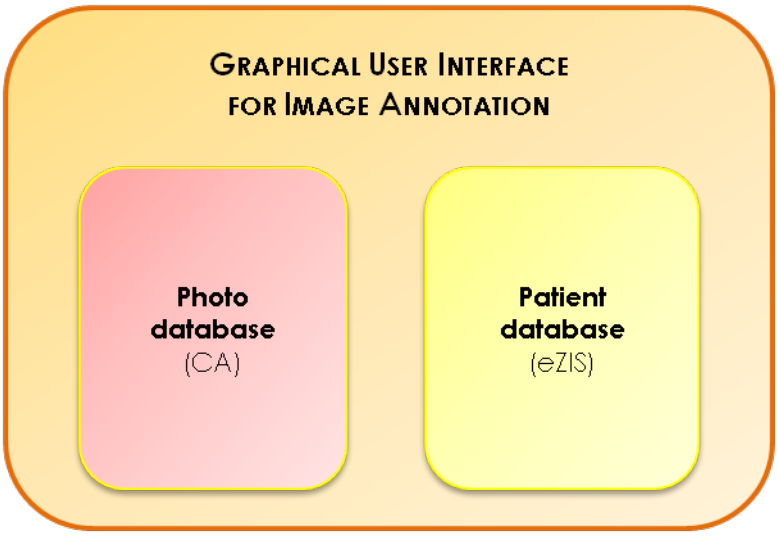


Figure 14
A GUI for image annotation

3.2 Step 2 – Semantic interoperability

When the basics of the previous step have been implemented into the medical practice, the achievement of semantic interoperability has been progressed. This paves the way to next step: benefits for the LUMC dermatology department, may or may not via other informatics techniques.

3.3 Step 3 – Benefits for the LUMC dermatology department

This section discusses example benefits for the LUMC dermatology department and assumes that the actions suggested in the previous sections have been implemented, thus that semantic interoperability has improved to a acceptable level. Also the current limitations of the informatics techniques will be considered as non-existent. Work to be done is described in section 4.3.

3.3.1 Content-based information retrieval

Properly annotated information, standard terminology and a good code system will result in the ability to retrieve information from a digital system based on its content.

Content-based retrieval can be further improved by IE, which is able to extract codes that are not explicitly stated from free text, and by integrating information. Another improvement could be the use of spell-checking during input, for instance using the UMLS. WordNet can be used to check for synonyms of non-medical words.

3.3.2 Gathering statistics

Improvements in content-based retrieval positively affect the abilities to gather statistics. Information Extractions allows integration of data, making it possible to gather statistics on data that previously have not been connected. Data Mining is able to detect interesting patterns that no one has thought of before, for instance sequences in disorders that are remarkable, allowing the hospital to tune its infrastructure to those sequences. Comparing these results with other institutions might yield interesting differences.

3.3.3 Computer-Aided Diagnosis & Telemedicine

The LUMC has expressed the ambition to provide new ways of healthcare to its patients, such as Telemedicine. ML/CV contribute to such progress in the dermatologic domain in the form of computer-aided diagnosis. Developments in mobile communication (e.g. smart phones), digital imaging and the Personal Health Record also contribute to the rise of telemedicine. A system like I^{2}Cnet[90], which is designed to discuss medical information over the internet, can also be used. For instance Kvedar *et al.* (1997)[91] showed that examining patients using dermatologic photographs, instead of directly, is possible.

4. CONCLUSIONS AND DISCUSSION

This chapter lists the recommendations and the techniques given in the previous chapters, discusses some of the limitations of this research, makes suggestions for further research and states the experiences of the author.

4.1 Conclusions

In order to benefit from Computer Science, the LUMC dermatology department needs to make changes in its infrastructure to achieve an acceptable level of semantic interoperability. These changes aim at implementing the usage of standard terminology and a code system, image annotation and standards for information exchange and imaging, into the medical practice. When semantic knowledge is integrated, information can be retrieved based on its content, there is an improved ability to gather statistics and computer-aided diagnosis and telemedicine can be performed. These all result in better healthcare, education and research.

Implementing any changes costs money but the benefits will eventually outweigh the costs.

Dermatology differs from other medical disciplines as their nomenclature is not as much integrated in code systems, and the DICOM standard is not used. Effort should be put into resolving this disparity.

All medical disciplines, computer scientists en ontologists should work together in designing systems that have optimised interoperability. Composing design rules for such systems is a good first step.

4.1.1 Achieving semantic interoperability

The LUMC dermatology department can improve the level of semantic interoperability by embracing the following recommendations:

Decent computer-processable nomenclature

- Recommendation 1: The LUMC should fully integrate SNOMED CT into its digital systems.
- Recommendation 2: While ICD-9 is used, effort should be put in fully integrating it into every IT system, including the DBC.
- Recommendation 3: While ICD-9 and CA are still used, the department can contact the makers of the applications to talk about less-rigorous steps than switching to a new code system.

Data annotation

- Recommendation 4: Researchers demand annotations of a higher quality than physicians do; to prevent unnecessary work the purpose of one annotation should be kept in mind.
- Recommendation 5: New image material should be annotated shortly after production. The usage of an old nomenclature like ICD-9, or the fact that data in the DBC are converted to a less advanced code, should be no excuse not to annotate
- Recommendation 6: Effort should be put in the annotation of old material.
- Recommendation 7: Physical annotation should be replaced by graphical annotation when possible.
- Recommendation 8: Image annotations should be in XML format.

- Recommendation 9: If possible, image annotation should map to ontologies like SNOMED CT and the FMA.
- Recommendation 10: As many information as possible should be annotated, such as age, sex, skin complexion and living place of the patient, diagnose, the primary body part, orientation of the body part, size of the photographed object, stage of the treatment and colour of the background.
- Recommendation 11: It should be kept in mind that non-doctors might need to work with the images, stretching the need to annotate whether the image should be classified for non-doctors.
- Recommendation 12: If the staff accidentally do not have enough time to annotate everything, at least the patient information, diagnose, skin complexion and primary body part should be annotated.
- Recommendation 13: A non-doctor such as a photographer should do the largest part of the annotation effort. The division of labour should be very clear. When that person its medical knowledge is insufficient, he or she can consult a physician.
- Recommendation 14: The person who annotates should be provided additional medical educated. This person is suggested to schedule the annotation-as-process at least weakly.
- Recommendation 15: Have medical students annotate images as part of their education, using a computer game.
- Recommendation 16: A classifier should be trained to give suggestions during the annotation process.
- Recommendation 17: Design a graphical interface that supports the annotation process. Allow the user to choose the input method of its choice.

Standards for information exchange and imaging

- Recommendation 18: Keep on using HL7.
- Recommendation 19: Orient on DICOM and prepare the systems for it. As soon as dermatological standards come available for it, the LUMC should start using it. As long as no guidelines for image annotation are available in DICOM, recommendation 5-8 should be followed.
- Recommendation 20: Design a graphical interface over CA and eZIS to facilitate the process of manual image annotation. While this interface is not present, the department secretary should stay responsible for transferring information between the two systems.

4.1.2 Benefits

Once semantic knowledge has been integrated, certain benefits can take place.

Content-based information retrieval can take place as a result of this data annotation. Information Extraction will be able to retrieve more information and integrate various systems. Spell-checking during input using the UMLS or WordNet will further increase the ability for content-based retrieval.

The reasons why content-based information retrieval can take place also apply to the improved ability to *gather statistics*, along with techniques for Information Extraction and Data Mining.

Computer-aided diagnosis and *telemedicine* are made possible by advancements in personal computing, smart phones and digital cameras, and by Computer Vision and Machine Learning.

4.1.3 Techniques

XML, *RDF*, *OWL*, *DL* and *Protégé-2000* laid the fundamental basis for semantic interoperability.

SNOMED CT, ICD, the *UMLS*, the *FMA* and *WordNet* make agreements about the meaning of terms thus contributing to the achievement of semantic interoperability. SNOMED is the most advanced code system.

Finally, *Health Level 7* and *DICOM* help in the achievement of semantic interoperability by making agreements on how data should look like. The *Integrating the Healthcare Enterprise* initiative stimulates the usage of HL7 and DICOM.

Machine Learning, *Computer Vision*, *Information Extraction* and *Data Mining* further contribute to achieving benefits.

Figure 15 summarises the steps to achieve benefits from computer science by achieving semantic interoperability.

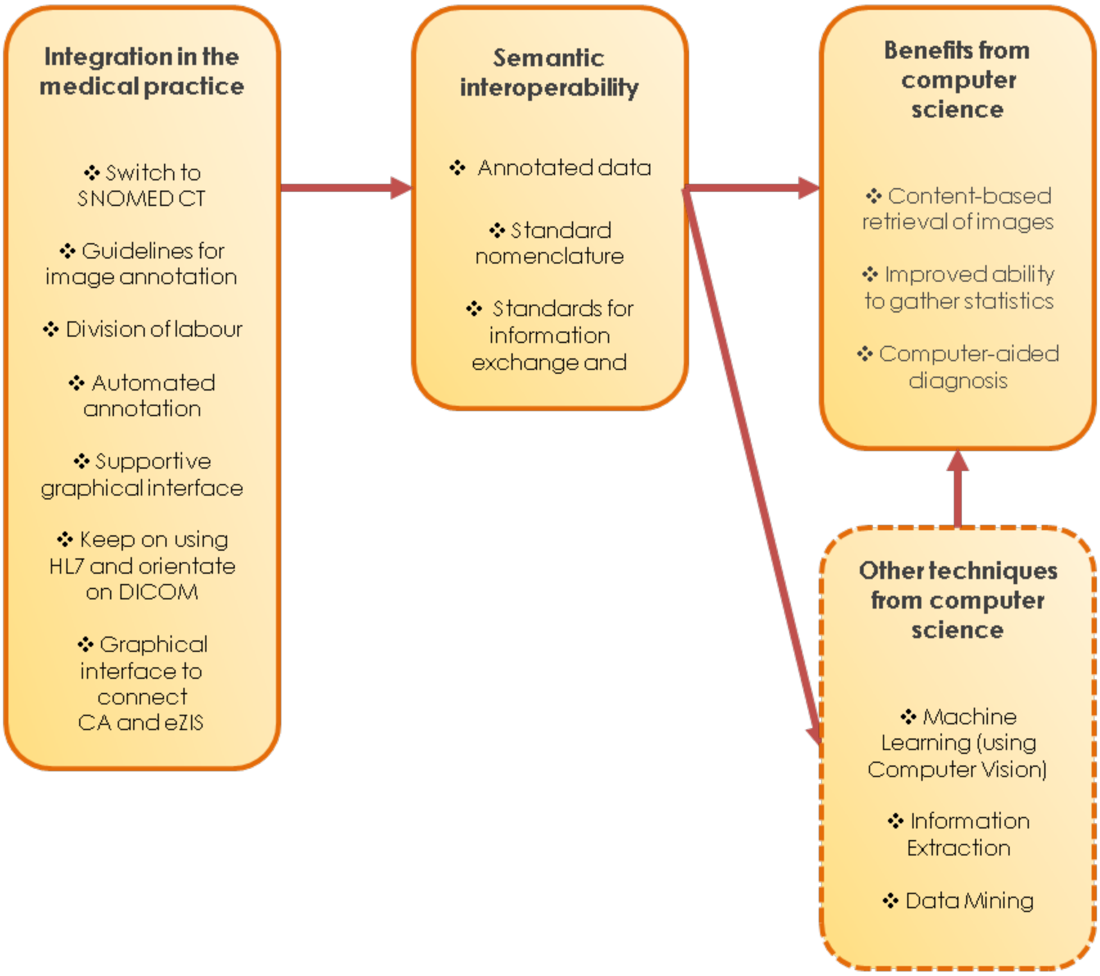


Figure 15
Summary of the steps to achieve benefits from computer science by achieving semantic interoperability.

4.2 Discussion

This thesis is based on a study of only a small part of the medical domain: the dermatology department of the LUMC hospital. Exploring the same objective for other disciplines in other, maybe non-academic, hospitals will likely yield different results.

When consulting individuals, more attention is needed in the problem formulation in such a way that the individuals are not offended. An extensive cost-benefit analysis helps supporting the point of switching to a new system. This thesis placed priority on the scientific aspect rather than on the consulting aspect, therefore little effort is put into a convincing rhetoric.

The process of implementing the recommendations should be workflow-driven: the implementers should be aware that big implementations require a major change of the information systems in the LUMC that will result in changes of the workflow of the medical staff. The key thus is to educate the staff and convince them that the necessary changes in their workflow will result in benefits for them. The first step is for the designers of improved or new systems: they have to take the current workflow of the staff as their starting point.

Also, the evolvement to new systems is costly and benefits will not directly be visible, and staff may be reluctant to change anything. Next to that have the current IT systems cost a lot of money; all of these reasons make the decision of evolving to new systems a very difficult one. Again, the solution is likely to show the staff what benefits will result from these changes.

No effort was put in defining what 'an acceptable level of semantic interoperability' really means, making it harder to determine what a good implementation is.

There always is an inconsistency between theoretical knowledge and the medical practice. Moreover, as computer science and medical science are two dissimilar disciplines there is an increased risk of misunderstanding during the implementation process.

Some techniques are not yet ready to be implemented so they need further research and development; they are described in section 4.3.

Finally, the opinion about the IT systems are collected from only two staff members; this is not a complete overview.

4.3 *Future work & recommendations*

This section describes what research and development needs to be done, and what needs to change in the current situation, in order for the desired benefits to take place.

DICOM

Dermatological standards need to be implemented into DICOM. DICOM needs to build in XML-based standards for annotation and it needs to be fully integrated in every medical discipline.

Terminology and code systems

All medical institutes in the world should eventually switch to the most advanced code system: SNOMED CT. The dermatological lexicon DermLex needs to be completed and implemented in SNOMED. Upon completion, it should contain all dermatological terminology, including the Dutch PROVOKE nomenclature.

Information Extraction and Data Mining

Work on IE needs to continue as no stable, consequent systems are available yet. Hospital data sets need to be made suitable for Data Mining.

Annotation game for students

Future research needs to point out whether a game that lets medical students annotate images is an effective way to decrease the effort of annotation from the medical staff.

Other ways to reduce annotation effort

Very little research has been done on ways how to minimise the annotation effort as part of the routine workflow of medical staff. Outcomes could be used both in- and outside the medical domain.

Computer-Aided Diagnosis

It is possible to discriminate between some skin disorders, but much more research is needed in this field of research.

Template matching for detecting limbs

A lot of research has been done on the detection of limbs in images with a macro view, but surprisingly little on images with a micro view. Template matching is a technique that has a high potential for success in detecting limbs in images with a micro view.

Skin complexion classification

Further research needs to be done on the automatic classification of skin complexions, into for instance classes in SNOMED.

4.4 Personal experiences

"I worked on the same project as Erwin Haas; I focussed on the general benefits of computer science whereas he covered the possibilities of the automatic classifications of skin diseases using Machine Learning and Computer Vision.

Finding a subject was difficult to me. The initial working title was "*Enriching a dermatology database using Computer Vision and Data Mining*", after which I switched to "*How to ensure that photos in a Dermatology Database are properly used*" because Computer Vision and Data Mining are not really my strong points. When typing the thesis, I first switched to "*Preventing data graveyards at the dermatology department of a large Dutch hospital, using image annotation and ontologies"* because it captures more of the computer science aspect; later I extended the subject to the possibilities of ML, CV and IE, ending up with the current research question.

I really enjoyed the consultancy aspect: how can a computer scientist help people with limited knowledge about the possibilities of informatics? Erwin and I visited the LUMC dermatology department several times and experienced that by listening carefully to their complaints and by asking questions, the core of the problem was revealed.

Erwin and I already helped some of the people at the department by setting up a website as part of a doctoral thesis: *www.africanskindiseases.org/wiki* (restricted and not yet finished; d.d. *28 September 2012*). The website is based on MediaWiki[s] and will eventually contain images of common skin diseases in Africa.

During the first months, most effort was put in building the website, experimenting with CV (the program *OpenCV*), choosing a subject, as well as other unrelated things such as organising a study trip; this resulted in a slow start-up. Eventually I just started to read theses after which many ideas popped up and started to write (hence the large number of references); after that everything went quite smoothly.

The staff at the LUMC was very helpful in providing all information that we needed. It is intriguing to see how a physician – who is very open towards informatics techniques and even built the internal dermatological extension on ICD – is being frustrated by the systems that are currently used.

After all, I am pleased with the results and I hope that the guidance of this thesis helps to improve the use of informatics techniques at the department."

Lucas van der Meer
23 August 2012

[s] *www.mediawiki.org*

AKNOWLEDGEMENTS

"I am very grateful to my supervisors, Fons J. Verbeek who guided me and provided me with a lot of feedback; my second reader Maarten Lamers; Sjan Lavrijsen and Thomas Donker from the Leiden University Medical Hospital who were very kind and helpful; my parents Hester and Reinier who carefully reviewed a draft version of this thesis; Arno Knobbe and Michel Chaudron who were very motivating; Peter Hendriks and Paul Kasteleyn who gave me useful suggestions; Erwin Haas with which I've worked on the same project; and all my friends who helped me through writing the thesis."

LITERATURE

[1] G. Eysenbach, "Consumer Health Informatics," *Bmj* 320, no. 7251 (2000): p1713–1716.

[2] Daniel Korenblum *et al.*, "Managing Biomedical Image Metadata for Search and Retrieval of Similar Images," *Journal of Digital Imaging* 24, no. 4 (2011): 739–748.

[3] N. Brown and M. Reynolds, *Strategy for Production and Maintenance of Standards for Interoperability Within and Between Service Departments and Other Healthcare Domains.* (Brussels, Belgium: Short Strategic Study CEN/TC, 2000).

[4] M. Eichelberg *et al.*, "A Survey and Analysis of Electronic Healthcare Record Standards," *ACM Computing Surveys (CSUR)* 37, no. 4 (2005): 277–315.

[5] B. Gibaud, "The Quest for Standards in Medical Imaging," *European Journal of Radiology* (2010).

[6] "WordNet Search - Information Technology", n.d., http://wordnetweb.princeton.edu/perl/webwn?s=information+technology.

[7] "LUMC Year Report 2011 (in Dutch)", n.d., http://www.lumc.nl/0000/12296/71203115213221/120322094854221/.

[8] "LUMC Dermatology Department Year Report 2009 (in Dutch)", n.d., 2009.

[9] Wikipedia contributors, "Dermatology," *Wikipedia, the Free Encyclopedia* (Wikimedia Foundation, Inc., July 4, 2012).

[10] "Wetten.nl - Wet- En Regelgeving - Zorgverzekeringswet - BWBR0018450", n.d., http://wetten.overheid.nl/BWBR0018450/geldigheidsdatum_25-07-2012.

[11] "Ben Ik Verplicht Een Zorgverzekering Af Te Sluiten? Rijksoverheid.nl", n.d., http://www.rijksoverheid.nl/onderwerpen/zorgverzekering/vraag-en-antwoord/ben-ik-verplicht-een-zorgverzekering-af-te-sluiten.html.

[12] Peter Hendriks, "Finding the Role of Disease-specific, Standardised Electronic Health Records in a Hospital's Information Architecture, Closing the Policy-Practise Gap; A Case-Study Approach" (2012).

[13] "Hospital Information Systems" (EHR Scope, n.d.), http://www.emrconsultant.com/education/hospital-information-systems.

[14] "Elektronisch Patiëntendossier Lijkt Te Sneuvelen in Senaat," *Nrc.nl*, n.d., http://www.nrc.nl/nieuws/2011/03/08/elektronisch-patientendossier-lijkt-te-sneuvelen-in-senaat/.

[15] R. H. Choplin *et al.*, "Picture Archiving and Communication Systems: An Overview.," *Radiographics* 12, no. 1 (1992): p127–129.

[16] B. C Madden, "A Proposal for Working Group 19–Dermatologic Standards" (n.d.), http://www.theskinappearancelaboratory.org/docs/DICOMwhitepaper.doc.

[17] R. Kumar, *Research Methodology: A Step-by-step Guide for Beginners* (Sage Publications Ltd, 2010).

[18] Tim Berners-Lee, Jamers Hendler, and Ora Lassila, "The Semantic Web," *Scientific American*, May 17, 2001.

[19] Michael K. Buckland, "Information as a Thing," *Journal of the American Society for Information Science* (1991): 351–360.

[20] W. J MacMullen, "Annotation as Process, Thing, and Knowledge: Multi-domain Studies of Structured Data Annotation," in *ASIST Annual Meeting*, 2005.

[21] Bryan Russell *et al.*, "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision* 77, no. 1 (2008): 157–173.

[22] D. L Rubin *et al.*, *Medical Imaging on the Semantic Web: Annotation and Image Markup (AIM)*, 2008.

[23] D. L Rubin and others, 'IPad: Semantic Annotation and Markup of Radiological Images', in AMIA Annual Symposium Proceedings, 2008, mmviii, 626 <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655990/> [accessed 19 June 2012].

[24] B. Hu *et al.*, "Ontology-based Medical Image Annotation with Description Logics," in *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference On*, 2003, 77–82.

[25] Alexander Maedche and Steffen Staab, *Learning Ontologies for the Sementic Web* (Institute AIFB, University of Karlsruhe, n.d.).

[26] W. Pidcock, "What Are the Differences Between a Vocabulary, a Taxonomy, a Thesaurus, an Ontology, and a Meta-model?" (2003).

[27] "W3C Mission", n.d., http://www.w3.org/Consortium/mission.html.

[28] Robin Cover, "XML Applications and Initiatives", n.d., http://xml.coverpages.org/xmlApplications.html.

[29] Wikipedia contributors, "XML," *Wikipedia, the Free Encyclopedia* (Wikimedia Foundation, Inc., July 12, 2012).

[30] Ora Lassila and Ralph R. Swick, *Resource Description Framework (RDF) Model and Syntax Specification*, n.d.

[31] Stefan Schulz *et al.*, "Strengths and Limitations of Formal Ontologies in the Biomedical Domain," *Revista Electronica De Comunicacao, Informacao; Inovacao EM Saude* 3, no. 1 (March 1, 2009): 31–45.

[32] D. L. McGuinness, F. Van Harmelen, and others, "OWL Web Ontology Language Overview," *W3C Recommendation* 10 (2004): 2004–03.

[33] I. Horrocks, "DAML+ OIL: a Reason-able Web Ontology Language," *Advances in Database Technology—EDBT 2002* (2002): 103–116.

[34] Natalya F. Noy *et al.*, "Protégé-2000: An Open-Source Ontology-Development and Knowledge-Acquisition Environment," *AMIA Annual Symposium Proceedings* 2003 (2003): 953.

[35] Ibid.; Holger Knublauch *et al.*, "The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications," in *The Semantic Web – ISWC 2004*, ed. Sheila McIlraith, Dimitris Plexousakis, and Frank van Harmelen, vol. 3298, Lecture Notes in Computer Science (Springer Berlin / Heidelberg, 2004), 229–243.

[36] "Nictiz.nl - Infrastructuur - Aorta (in Dutch)", n.d., www.nictiz.nl/page/Infrastructuur.

[37] Eliot L. Siegel and David S. Channin, "Integrating the Healthcare Enterprise: A Primer Part 1. Introduction1," Radiographics 21, no. 5 (September 1, 2001): 1339–1341.

38 A. Gangemi, D. M Pisanelli, and G. Steve, "An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies," Data & Knowledge Engineering 31, no. 2 (1999): 183–220.

[39] "International League of Dermatologic Societies - ICD-10", n.d., http://web.ilds.org/cms/index.php?page=about-icd-10.

[40] Wikipedia contributors, "SNOMED CT," *Wikipedia, the Free Encyclopedia* (Wikimedia Foundation, Inc., July 3, 2012).

[41] "Nictiz.nl - SNOMED CT (in Dutch)", n.d., http://www.nictiz.nl/page/Standaarden/SNOMED-CT.

[42] *Stageboek Dermatologie (in Dutch)* (Leiden: LUMC, 2010).

[43] T. M. Starink, J. J. E. Van Everdingen, and M. de Haan, *Dermatovenereologie Voor De 1e Lijn* (Bohn Stafleu van Loghum, 2009), 21–27.

[44] A. Papier *et al.*, "Framework for Improved Communication: The Dermatology Lexicon Project," *Journal of the American Academy of Dermatology* 50, no. 4 (2004): 630–634.

[45] Alicia Ault, "Dermatology Lexicon Web Site to Be Launched by AAD," *Skin and Allergy News*, March 2009, 40 issue 3 edition.

[46] B. R. Eapen, "ONTODerm–a Domain Ontology for Dermatology.," *Dermatology Online Journal* 14, no. 6 (2008): 16.

[47] J. W Woods *et al.*, "Using UMLS Metathesaurus Concepts to Describe Medical Images: Dermatology Vocabulary," *Computers in Biology and Medicine* 36, no. 1 (2006): 89–100.

[48] H. Tolentino *et al.*, "A UMLS-based Spell Checker for Natural Language Processing in Vaccine Safety," *BMC Medical Informatics and Decision Making* 7, no. 1 (2007): 3.

[49] C. Rosse, J. L.V Mejino, and others, "A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy," *Journal of Biomedical Informatics* 36, no. 6 (2003): 478–500.

[50] O. Bodenreider and A. Burgun, "Biomedical Ontologies," *Medical Informatics* (2005): 211–236.

[51] Olivier Bodenreider and Anita Burgun, "Characterizing the Definitions of Anatomical Concepts in WordNet and Specialized Sources", n.d.

[52] Li Zhang *et al.*, "Enriching the Structure of the UMLS Semantic Network", n.d.

[53] George A. Miller, "WordNet: a Lexical Database for English", n.d.

[54] Jorge Morato *et al.*, "WordNet Applications", n.d.

[55] Wikipedia contributors, "WordNet," *Wikipedia, the Free Encyclopedia* (Wikimedia Foundation, Inc., June 10, 2012).

[56] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic Image Annotation and Retrieval Using Cross-media Relevance Models," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 2003, 119–126.

[57] P. Duygulu *et al.*, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," *Computer Vision—ECCV 2002* (2006): 349–354.

[58] J. Yao *et al.*, "Automatic Medical Image Annotation and Retrieval Using SECC," in *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium On*, 2006, 820–825.

[59] J. Kalpathy-Cramer and W. Hersh, "Automatic Image Modality Based Classification and Annotation to Improve Medical Image Retrieval," *Studies in Health Technology and Informatics* 129, no. 2 (2007): 1334.

[60] A. Mueen, R. Zainuddin, and M. S Baba, "Automatic Multilevel Medical Image Annotation and Retrieval," *Journal of Digital Imaging* 21, no. 3 (2008): 290–295.

[61] H. Müller *et al.*, "A Review of Content-based Image Retrieval Systems in Medical Applications—clinical Benefits and Future Directions," *International Journal of Medical Informatics* 73, no. 1 (2004): 1–23.

[62] Erwin Haas, "Classifying Dermatological Diseases Using Computer Vision (unpublished Work)", 2012.

[63] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A Survey on Pixel-based Skin Color Detection Techniques," in *Proc. Graphicon*, vol. 3, 2003.

[64] J. Y. Lee and S. I. Yoo, "An Elliptical Boundary Model for Skin Color Detection," in *Proc. of the 2002 International Conference on Imaging Science, Systems, and Technology*, 2002, http://agent.snu.ac.kr/papers/IC/ic_cisst_2002_elliptical_boundary_jylee.pdf.

[65] T. M. Mahmoud, "A New Fast Skin Color Detection Technique," *World Academy of Science* (2008): 501–505.

[66] D. Brown, I. Craw, and J. Lewthwaite, "A Som Based Approach to Skin Detection with Application in Real Time Systems," in *Proc. of the British Machine Vision Conference*, vol. 2, 2001, 491–500.

[67] B. D. Zarit, B. J. Super, and F. K.H. Quek, "Comparison of Five Color Models in Skin Pixel Classification," in *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999. Proceedings. International Workshop On*, 1999, 58–63, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=799224.

[68] J. Brand and J. S. Mason, "A Comparative Assessment of Three Approaches to Pixel-level Human Skin-detection," in *Pattern Recognition, 2000. Proceedings. 15th International Conference On*, vol. 1, 2000, 1056–1059.

[69] B. Jedynak, H. Zheng, and M. Daoudi, "Maximum Entropy Models for Skin Detection," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2003, 180–193.

[70] G. Schaefer *et al.*, "Colour and Contrast Enhancement for Improved Skin Lesion Segmentation," *Computerized Medical Imaging and Graphics* 35, no. 2 (2011): 99–104.

[71] M. E. Celebi *et al.*, "Lesion Border Detection in Dermoscopy Images," *Computerized Medical Imaging and Graphics* 33, no. 2 (2009): 148–153.

[72] "Computer-aided Diagnosis - Wikipedia, the Free Encyclopedia", n.d.

[73] P. Schmid-Saugeona, J. Guillodb, and J. P. Thirana, "Towards a Computer-aided Diagnosis System for Pigmented Skin Lesions," *Computerized Medical Imaging and Graphics* 27, no. 1 (2003): 65–78.

[74] T. Schindewolf *et al.*, "Evaluation of Different Image Acquisition Techniques for a Computer Vision System in the Diagnosis of Malignant Melanoma," *Journal of the American Academy of Dermatology* 31, no. 1 (1994): 33–41.

[75] A. Dalal *et al.*, "Concentric Decile Segmentation of White and Hypopigmented Areas in Dermoscopy Images of Skin Lesions Allows Discrimination of Malignant Melanoma," *Computerized Medical Imaging and Graphics* 35, no. 2 (2011): 148–154.

[76] M. Emre Celebi *et al.*, "Automatic Detection of Blue-white Veil and Related Structures in Dermoscopy Images," *Computerized Medical Imaging and Graphics* 32, no. 8 (December 2008): 670–677.

[77] I. Haritaoglu, D. Harwood, and L. S. Davis, "Ghost: A Human Body Part Labeling System Using Silhouettes," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference On*, vol. 1, 1998, 77–82.

[78] I. Haritaoglu, D. Beymer, and M. Flickner, "Ghost3D: Detecting Body Posture and Parts Using Stereo", 2002, 175–180.

[79] I. Haritaoglu, D. Harwood, and L. Davis, "W 4 S: A Real-time System for Detecting and Tracking People in 2 1/2D," *Computer Vision—ECCV'98* (1998): 877–892.

[80] R. Polana and R. Nelson, "Low Level Recognition of Human Motion (or How to Get Your Man Without Finding His Body Parts)," in *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop On*, 1994, 77–82.

[81] L. S Overney *et al.*, "Cerebral Processes in Mental Transformations of Body Parts: Recognition Prior to Rotation," *Cognitive Brain Research* 25, no. 3 (2005): 722–734.

[82] T. Deselaers *et al.*, "The CLEF 2005 Automatic Medical Image Annotation Task," *International Journal of Computer Vision* 74, no. 1 (2007): 51–58.

[83] T. Tommasi *et al.*, "Overview of the CLEF 2009 Medical Image Annotation Track," *Multilingual Information Access Evaluation II. Multimedia Experiments* (2010): 85–93.

[84] M. A. Hearst, "Untangling Text Data Mining," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, 3–10.

[85] G. DeJong, "An Overview of the FRUMP System," *Strategies for Natural Language Processing* (1982): p149–176.

[86] S. M. Meystre *et al.*, "Extracting Information from Textual Documents in the Electronic Health Record: a Review of Recent Research," *Yearb Med Inform* 35 (2008): p128–144.

[87] "2007 International Challenge: Classifying Clinical Free Text Using Natural Language Processing | Computational Medicine Center", n.d., http://computationalmedicine.org/challenge/previous.

[88] N. Ramakrishnan, D. Hanauer, and B. Keller, "Mining Electronic Health Records," *Computer* 43, no. 10 (2010): 77–81.

[89] L. Von Ahn and L. Dabbish, "Labeling Images with a Computer Game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2004, 319–326.

[90] C. E. Chronaki, X. Zabulis, and S. C. Orphanoudakis, "I2Cnet Medical Image Annotation Service," Informatics for Health and Social Care 22, no. 4 (1997): 337–347.

[91] J. C. Kvedar *et al.*, "The Substitution of Digital Images for Dermatologic Physical Examination," *Archives of Dermatology* 133, no. 2 (February 1, 1997): p161–167.a

GLOSSARY

CA	Clinical Assistant, the PACS used at the LUMC
CAD	Computer-Aided Diagnosis
CV	Computer Vision
DBC (Casemix)	Diagnose-Behandelcombinatie (*Casemix*); a system to encourage efficiency in the healthcare system.
DICOM	Digital Imaging and Communications in Medicine; a standard for storage and exchange of medical images.
DL	Description Logic
EHR (EPD)	Electronic Health Record (*Electronisch Patiëntendossier*)
FMA	Foundational Model of Anatomy
HCI	Human-computer interaction. The way that humans and computers interact with each other, and how humans judge this interaction.
HIS (ZIS)	Hospital Information System (*ZiekenhuisInformatieSysteem*)
HL7	Health Level 7; a standard used for information exchange in the medical domain
ICD	International Statistical Classification of Diseases and Related Health Problems
IE	Information Extraction; an informatics technique
IHE	Integrating the Healthcare Enterprise; an initiative to promote the usage of standards such as DICOM and HL7.
LUMC	Leiden University Medical Centre (*Leids Universitair Medisch Centrum*)
ML	Machine Learning; an informatics technique
Nictiz	National IT Institute for Healthcare in the Netherlands (*Nederlands Instituut voor ICT in de Zorg*)
OWL	Web Ontology Language
PACS	Picture Archiving and Communication System
PHR	Personal Health Record
PROVOKE	Dutch system used to describe skin disorders
RDF	Resource Description Framework
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
UMLS	Unified Medical Language System
XML	Extensible Markup Language